DYNAMIC CONTROL CHARTS: A DIAGNOSTIC TEST FOR SUBSTANTIATING IMPROVEMENT IN EDUCATION

by
Nathaniel Andrews Dewey

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
July 2021

Abstract

Continuous improvement – a well-known strategy for iteratively advancing quality in industry and health care – has garnered substantial support in the education sector over the past decade. One of the distinguishing features of continuous improvement is the use of a disciplined methodology, e.g., plan-do-study-act cycles, to rapidly ideate, implement, and test change ideas. The ability to quickly provide a strong warrant that a change idea led to improvement is integral to the success of this approach. In industry, the primary method for providing this warrant is statistical process control (SPC) – a set of statistical diagnostic tests designed to be used by engineers in monitoring simple machine processes. Unfortunately, in education, processes are often complicated by human social dynamics making traditional SPC unworkable. Furthermore, traditional SPC demands a statistical skill set that is less prevalent among educators. Given these difficulties, an alternative approach is needed. In this dissertation, I introduce the Dynamic Control Chart – a novel diagnostic test for substantiating improvement based on statistical techniques designed for making causal claims using automated short-term forecasts.

The primary aims of this dissertation are to determine the methodological requirements for substantiating improvement in education and to test the adequacy of Dynamic Control Charts in meeting these requirements. I first identify five requirements for substantiating improvement based on the existing strengths of statistical process control and the relevant differences between industry and education. Specifically, I find the method needs to be disciplined but pragmatic, appropriate for small samples, responsive with limited data,

semiparametric, and unobtrusive and automated. I then conduct simulation studies to test the adequacy of Dynamic Control Charts in meeting two of the requirements, namely that the method be responsive with limited data and semiparametric. Using simulations, I find that the Dynamic Control Chart provides stronger confirmatory tests than traditional SPC methods making it responsive with limited data. I also find that the Dynamic Control Chart has the tools to accommodate different forms of temporal dynamicity and in monthly attendance data performs excellently by the standards of the educational literature.

Primary Reader and Advisor: Marc L. Stein

Acknowledgements

I am incredibly grateful to my advisor, Professor Marc Stein, who introduced me to improvement science and connected me to the burgeoning educational community in the school improvement space. When I first began working with Marc, my personal research agenda was entirely different and my enthusiasm for educational research was waning. Marc's excitement about his work and the potential he saw in me changed everything. In Marc's wake, I found a new research agenda which was more in line with my talents and inclinations, and I found my purpose in educational research. Without Marc's guidance and support, I do not know that I would have made it to this moment.

I am also grateful to Prof. Julia Burdick-Will, Prof. Faith Connolly and Prof. Robert Balfanz for serving on my various dissertation committees, and Prof. Odis Johnson and Prof. Anthony Bryk for serving on my Graduate Board Oral committee.

Dedication

This dissertation is dedicated to my wife, Leah Rose Dewey, who encouraged me to pursue my doctorate and then started a business to support our family while I went back to school. This all would have been impossible without her love and trust.

Contents

Abstract ii

Acknowledgements iv

Dedication v

List of Tables ix

List of Figures x

Chapter 1 – Foreword 1

Chapter 2 – Substantiating Improvement in Education: The Challenge of Translating Methods from Industry 6

Introduction 6

Methods for Substantiating Improvement in Industry and Health Care 14

Control Charts 16

Run Charts 22

Differences between Education and Other Sectors 25

Subjects of Improvement 26

Data and Measurement 28

Practitioners and the Organization 30

Research and Development 31

Methodological Requirements for Substantiating Improvement in Education 33

Disciplined but pragmatic 37

Appropriate for Small Samples of Subjects 39

Responsive with Limited Time-series Data 41

Semiparametric 43

Unobtrusive and Automated 45

Discussion 47

Chapter 3 – Statistical Process Control using Bayesian Structural Time-Series Models 51

Introduction 51

Methods 58

Simulation of Data 59

Tested Methods 61

Likelihood Ratios 72

Results 73

Tests on Exchangeable Data 75

Tests on Autocorrelated Data 77

Discussion 78

Conclusion 81

Chapter 4 – Receiver Operating Characteristic Analysis of Dynamic Control Charts: A Case Study of Monthly School Attendance 83

Introduction 83

Methods 86

Daily Attendance Data 87

Simulation of Attendance Effects 91

Dynamic Control Charts 93

Receiver Operating Characteristic (ROC) Curve Analysis 98

Results 100

Dynamic control chart with a static intercept (1) 101

Dynamic control chart with a local level component (2) 103

Dynamic control chart including the school's previous year's attendance as a predictor (3) 104

Dynamic control chart including the mean contemporaneous attendance of similar schools as a predictor (4) 105

Discussion 106

Chapter 5 – Afterword 110

Conclusion 113

References 117

List of Tables

Table 2.1. Comparison Between Methodological Needs for CI in Education and Various Methods from Educational Research and Statistical Process Control 49

List of Figures

Figure 2.1. An example of a Shewhart X-bar chart where non-random change is present. The black line graph is the mean outcome data for the samples (N=10, M=0, SD=1). The dashed vertical line is the beginning of the non-random mean shift (D=1 SD). The solid horizontal line (CL) is the mean of the calibration data. The dashed horizontal lines (UCL & LCL) are the control limits for an X-bar chart (3 standard deviations). The red and orange dots in the new data are out of control signals. 18

Figure 2.2. An example of a EWMA chart where non-random change is present. The black line graph is the exponentially weighted moving average of the mean outcome data for the samples (N=10, M=0, SD=1). The dashed vertical line is the beginning of the non-random mean shift (D=1 SD). The solid horizontal line (CL) is the mean of the calibration data. The dashed horizontal lines (UCL & LCL) are the control limits for an EWMA chart (3 standard deviations). The red dots in the new data are out of control signals. 20

Figure 2.3. An example of a Run Chart where non-random change is present. The blue line graph is the raw outcome data (N=10, M=0, SD=1). The black dotted vertical line is the beginning of the non-random mean shift (D=1 SD). The red dashed horizontal line is the median of the calibration data. 23

Figure 3.1. An example of a Run Chart where non-random change was present and detected by Anhøj decision rules. The blue line graph is the raw outcome data (M=0, SD=1). The dashed line is the beginning of the non-random mean shift (D=2). The grey line is the median of the pre period. 63

Figure 3.2. An example of a Shewhart I-chart where non-random change was present and detected by WE decision rules 1-3. The blue line graph is the raw outcome data (M=0, SD=1). The dashed line is the beginning of the non-random mean shift (D=2). The grey line is the mean of the pre period. The grey ribbon is the region bounded by the 3-sigma control limits. 65

Figure 3.3. An example of a Dynamic Control Chart Intercept Model where non-random change was present and detected (p=0.0011). The blue line graph is the raw outcome data (M=0, SD=1). The dashed line is the beginning of the non-random mean shift (D=2). The grey line is the model fit to the left of the dashed line, and the counterfactual to the right of the dashed line. The grey ribbon is the 95% Bayesian credible interval. 69

Figure 3.4. An example of a Dynamic Control Chart Autoregressive Model where non-random change was present and detected (p=0.01). The blue line graph is the raw outcome data (M=0, SD=1, Ar1=0.4). The dashed line is the beginning of the non-random mean shift (D=2). The grey line is the model fit to the left of the dashed line, and the counterfactual to the right of the dashed line. The grey ribbon is the 95% Bayesian credible interval..... 71

Figure 3.5. Likelihood ratios for SPC methods given normally distributed exchangeable data (M=0, SD=1) with various pre and post period lengths. 74

Figure 3.6. Likelihood ratios for SPC methods given autocorrelated data with normally distributed innovations (M=0, SD=1) and with various post period lengths and levels of autoregression. 75

Figure 4.1. Example of an outlier in daily attendance data (red) which was discovered to be a half-day for parent teacher conferences. 89

Figure 4.2. Example of a weekly cycle in daily attendance data where attendance in the middle of the week is higher than attendance at the ends of the week. Each group of five columns is a single week. The blue line is a loess curve selected to highlight the cycle. 89

Figure 4.3. Example of localized trending present in weekly attendance data. From week 1 to 9 attendance is on a downward trend and from week 10 to 15 attendance is on an upward trend. The blue line is a loess curve selected to highlight the trending. 90

Figure 4.4. Example of seasonality in monthly attendance data. The blue line is a loess curve selected to highlight the seasonality. 90

Figure 4.5. Examples of four types of dynamic control charts where non-random change was present and detected. The blue line graph is the monthly average attendance percentage data. The dashed line is the beginning of the non-random mean shift (+5). The grey line is the model fit to the left of the dashed line, and the counterfactual to the right of the dashed line. The grey ribbon is the 95% Bayesian credible interval. 4.5a. Dynamic control chart with a static intercept. 4.5b. Dynamic control chart with a local level component. 4.5c. Dynamic control chart including the school's previous year's attendance as a predictor. 4.5d. Dynamic control chart including the mean contemporaneous attendance of similar schools as a predictor. 95

Figure 4.6. Example of a symmetrical ROC curve with an AUC of 0.90. 99

Figure 4.7. ROC curve of dynamic control chart with a static intercept for various mean shift effects (+1%, +5%, and +10%). 102

Figure 4.8. Examples of four types of dynamic control charts applied to unaltered data with dynamicity. The blue line graph is the monthly average attendance percentage data. The dashed line is the end of the calibration. The grey line is the model fit to the left of the dashed line, and the counterfactual to the right of the dashed line. The grey ribbon is the 95% Bayesian credible interval. 4.8a. Dynamic control chart with a static intercept. 4.8b. Dynamic control chart with a local level component. 4.8c. Dynamic control chart including the school's previous year's attendance as a predictor. 4.8d. Dynamic control chart including the mean contemporaneous attendance of similar schools as a predictor. 103

Figure 4.9. ROC curve of dynamic control chart with a local level component for various mean shift effects (+1%, +5%, and +10%). 104

Figure 4.10. ROC curve of a dynamic control chart including the school's previous year's attendance as a predictor for various mean shift effects (+1%, +5%, and +10%). 105

Figure 4.11. ROC curve of dynamic control chart including the mean contemporaneous attendance of similar schools as a predictor for various mean shift effects (+1%, +5%, and +10%). 106

Figure 4.12. ROC curves of four types of dynamic control charts for a mean shift of +5%. 108

Chapter 1 – Foreword

At the dawn of the 21st century, educational research and reform was undergoing a promising transformation. Decades of increasing school accountability coupled with frustration over the failures of educational reform efforts had generated a policy environment primed for change. This opportunity was taken up by the evidence-based reform movement which sought to increase the rigor of educational research by encouraging experimentation in education. Remarkable advances during the previous century in medicine, agriculture, and other fields were attributed to the use of randomized controlled trials. Educational scholars in the evidence-based reform movement argued that an increase in experimentation driven by evidence-based policies could realize a scientific revolution in education (Slavin, 2002).

Over the next two decades evidence-based reformers saw many policy victories including the creation of the Institute for Education Sciences (IES) in 2002, the subsequent establishment of rigorous research training programs (e.g., 2004's Predoctoral Interdisciplinary Research Training Program in the Education Sciences) and evidence-based research grant programs (e.g., 2009's Investing in Innovation), and more recently, the 2015 passage of the Every Student Succeeds Act (ESSA) which defined tiers of evidence for educational programs and tied some forms of federal funding for education to the use of evidence-based programs. The policy changes over this period reshaped the landscape of educational research. Since 2004, over 1000 researchers have been trained to conduct rigorous experiments in education, and $1.4 billion in grant money has been spent to develop, evaluate, and scale-up evidence-based programs. Furthermore, as hoped for by evidence-based reformers, the number of

experimental and quasi-experimental studies conducted each year in education has increased markedly since the turn of the century (Slavin, 2020).

Regrettably, over this same period there have not been simultaneous gains in measures of educational progress. Evidence-based reformers have suggested that the sluggish gains can be attributed to the disconnect between research and practice (Slavin, 2020). However, a diverse group of educational scholars outside of the evidence-based movement have argued from multiple positions that the problem is deeper (see G. J. Biesta, 2010; Bryk et al., 2015; Deaton & Cartwright, 2018; Elmore, 2016). Evidence-based reform has underestimated the highly contextual nature of education, placed too much confidence in the randomized controlled trial, and consequently, struggled to spread and scale 'proven' programs because the evidence is often not generalizable, potentially even in the same school in the next school year (for an example see Hanselman et al., 2017). This should be somewhat unsurprising given decades of retrospective scholarship on educational reform (see Mehta (2015) for accountability; Honig (2006) for implementation; Hess (2011) and Tyack & Cuban (1995) for reform) has advocated for a greater role for local knowledge and actors (i.e., contextual variation) in educational initiatives.

Continuous improvement – a widespread strategy for advancing quality in industry (Deming, 2000; Langley et al., 2009) and health care (Schouten et al., 2008) – has arisen as a new avenue for educational reform which is scientific and disciplined, much like evidence-based reform, but which explicitly attends to contextual variation through a focus on learning from small adjustments using an iterative process (Bryk et al., 2015; Yurkofsky et al., 2020). In 2018, sixteen states acknowledged continuous improvement as an essential aspect of their

ESSA theory of action (Results for America, 2018) and many more chose to include the language of continuous improvement in their implementation plans or already had continuous improvement embedded in their state education policy (Hough et al., 2017). Education researchers have also recently become involved in continuous improvement through federal grants supporting the construction of research-practice partnerships (Cohen-Vogel et al., 2014) and large investments from private funders including the Carnegie Foundation (Improvement Science) and the Bill & Melinda Gates Foundation (Networks for School Improvement). Consequently, continuous improvement will likely influence education nationwide in the coming years. Though, whether continuous improvement ultimately leads to gains in measures of educational progress remains to be seen, especially in urban districts where reforms frequently falter (Payne, 2008).

One of the distinguishing features of continuous improvement is the use of a disciplined methodology, e.g., plan-do-study-act (PDSA) cycles, to rapidly ideate, implement, and test change ideas (Bryk et al., 2015; Deming, 2000; Hough et al., 2017). The ability to quickly provide a strong warrant that a change idea led to improvement is integral to the success of this approach (Reed & Card, 2016). In industry, the primary method for providing this warrant is statistical process control (SPC) – a set of statistical diagnostic tests designed to be used by engineers in monitoring machine processes (Shewhart, 1931; Langley et al., 2009). Unfortunately, in education, processes are complicated by human social dynamics (Koopmans, 2020; National Research Council, 2002) and data is created and collected on a much slower cadence potentially making traditional SPC unworkable. Furthermore, traditional SPC demands a statistical skill set that is less prevalent among educators (Chirume, 2018; Roderick, 2012).

Given these difficulties, an alternative approach for substantiating improvement is needed to realize the benefits of continuous improvement in education.

Recently, with the proliferation of big data, methodologies for complex automated analyses have grown in popularity. These methodologies address analytic decision making either empirically or heuristically to reduce the cost associated with conducting thousands of analyses directly. For instance, automated short-term forecasting, also called nowcasting, has been used by online advertising providers to present clients with timely impact estimates on advertising campaigns (Brodersen et al., 2015; Scott & Varian, 2014). In the present dissertation, I introduce the Dynamic Control Chart – a novel diagnostic test for substantiating improvement in education that is based on these statistical techniques for making causal claims using automated short-term forecasts. The Dynamic Control Chart is presented as a simple plot with an accompanying diagnostic test available in an R package written for this dissertation (https://github.com/westdew/dccharts/). Behind this facade, however, is an automated methodology based on the R package Causal Impact (Brodersen et al., 2015) which fuses state-space models (Durbin & Koopman, 2012) and Bayesian automatic variable selection (George & McCulloch, 1997) from applied statistics with synthetic control methods from political science's quantitative comparative case-studies (Abadie et al., 2015) to make causal claims using automated short-term forecasts (Scott & Varian, 2014).

The goals of the present dissertation, then, are to further motivate the need for Dynamic Control Charts and test the adequacy of Dynamic Control Charts as a diagnostic test for substantiating improvement in education. To meet these goals, I will address the following three aims:

Aim 1. Determine the methodological requirements for substantiating improvement in education.

Aim 2. Compare the performance of Dynamic Control Charts and traditional statistical process control methods for simulated data.

Aim 3. Examine the accuracy of Dynamic Control Charts for a real-world educational data context, namely attendance data with a simulated effect.

The present dissertation consists of three academic studies (Chapters 2-4) each addressing one of the dissertation's aims. Together these studies will show the Dynamic Control Chart to be a necessary and adequate methodological innovation in educational continuous improvement work. Furthermore, these studies will demonstrate the potential of the Dynamic Control Chart to address a real problem in education, namely how do educators determine if a change they made was an improvement. The rise of continuous improvement could realize a decades long shift in reform efforts towards an emphasis on local knowledge and actors, but it may struggle to do so without the ability to answer this perennial question using disciplined scientific methods.

Chapter 2 – Substantiating Improvement in Education: The Challenge of Translating Methods from Industry

Introduction

Continuous improvement – a well-known strategy for gradually advancing quality in industry (Deming, 2000) and health care (Provost & Murray, 2011) – has garnered substantial support in the education sector over the past decade (Grunow et al., 2018). Most recently, sixteen states acknowledged continuous improvement as an essential component of their Every Student Succeeds Act (ESSA) theory of action (Results for America, 2018) and many more states included the language of continuous improvement in their ESSA implementation plans or already had elements of continuous improvement embedded in their state education policy. California, for instance, did not name continuous improvement as essential to their ESSA theory of action yet have supported continuous improvement in their schools for over a decade now (Hough et al., 2017).

Education researchers have also become increasingly involved in continuous improvement work in this period through federal grants that supported the construction of research-practice partnerships (Cohen-Vogel et al., 2014) and a growing dissatisfaction and concern with solution-based (i.e., 'what works') education reform (see G. J. Biesta, 2010; Bryk et al., 2015; Deaton & Cartwright, 2018; Elmore, 2016). Additionally, large investments from non-profit and philanthropic organizations including the Carnegie Foundation (Improvement Science) and the Bill & Melinda Gates Foundation (Networks for School Improvement) may

have attracted researchers to improvement work. As a consequence of this coalescence of political, academic, and monetary support, continuous improvement will likely have an effect on education nationwide in the coming years. Whether that effect will translate into better student outcomes is uncertain, especially in urban districts where reforms frequently falter (Payne, 2008).

One reason it is difficult to imagine how continuous improvement will affect education is the ambiguity of the term. In a chapter examining the set of ideas and practices in education now collectively referred to as continuous improvement, Yurkofsky et al. (2020) found there were at least 14 distinct continuous improvement methods in the education literature. In part this is because continuous improvement has an inherent, clearly understood, and unobjectionable meaning even if the technical details of the practice as defined and used by W. Edwards Deming, the father of quality improvement in industry (see Deming, 2000), are not familiar. Furthermore, as evidenced by the volume of distinct methods under the same name (Yurkofsky et al., 2020), continuous improvement may have taken on many new meanings and, as a term, been applied to other distinct educational research paradigms (e.g., research-practice partnerships) as it has been translated from health care and industry to education (see Cohen-Vogel et al., 2014).

Even though continuous improvement does not have a single definition across educational stakeholders (Hough et al., 2017), a consensus has still emerged around the importance of continuous improvement in education, suggesting that these ideas and practices share distinguishing features that separate them from other educational initiatives. Although there are some differences between recent attempts to articulate the commonalities between

continuous improvement methods (e.g., Hough et al., 2017; and Yurkofsky et al., 2020), scholars have largely identified similar sets of shared characteristics. The characteristics include: 1) a focus on local problems and needs, 2) a perspective that systems contribute to outcomes more than individuals, 3) a commitment to empowering educators to engage in improvement work, and 4) a belief in the use of disciplined iterative methodologies to solve problems. These characteristics stand out mostly for their contrast with those of past educational initiatives, e.g., school accountability, implementation science, and comprehensive school reform. Whereas the distinguishing features of continuous improvement engage and empower educators by placing the onus on the system educators work within, past educational initiatives have been accused of doing the opposite. Retrospective scholarship (see Mehta (2015) for accountability; Honig (2006) for implementation; Hess (2011) and Tyack & Cuban (1995) for reform) has advocated for a greater role for local knowledge and actors in educational initiatives for decades. Remarkably, continuous improvement has the potential to realize this transformation.

With rising interest in continuous improvement across the education sector, the challenges of applying existing improvement methods from industry to educational problems have begun to be explored by "critical friend[s] of the movement" (Yurkofsky et al., 2020, p. 404). For example, Yurkofsky et al. (2020) argue continuous improvement methods need to be more attentive to the relational and political dimensions of improvement in educational contexts. Otherwise, continuous improvement may become "the myth and ceremony of the modern age" of education (Yurkofsky et al., 2020, p. 425). The present study takes the same benevolent stance in its criticism of current continuous improvement methods. However, instead of focusing a critical lens on the methods as a whole, I focus on one narrow but crucial

aspect of continuous improvement that I contend has been lost in translation from industry and health care – the challenge of substantiating that a change is an improvement.

Continuous improvement has gained much if its support in education by hearkening back to successes from industry and health care (Bryk et al., 2015) which were based on the science of improvement (Deming, 2000; Shewhart, 1931). A critical component of continuous improvement is the use of rigorous, scientifically grounded methods for testing changes during the work. These disciplined methods provide much of the evidentiary support for continuous improvement's ability to lead to meaningful change. Education though is fundamentally different from industry and health care (National Research Council, 2002). The methods used to substantiate improvement in these sectors may be difficult to translate to improvement work in schools. Researchers in the field of education can already attest to the challenges of experimentation in educational contexts. The same challenges will be present when translating improvement methods.

Although there is no common standard for iteratively testing changes in educational continuous improvement work, the most well-known method is the Plan-Do-Study-Act (PDSA) cycle (Bryk et al., 2015; Langley et al., 2009).[1] A PDSA cycle is a rapid microcosmic version of the scientific method whereby change ideas are hypothesized (Plan), enacted (Do), and tested (Study), before being fed back into the next iteration of development (Act). In the 'Study' phase of a PDSA cycle, the data from the 'Do' phase is analyzed to determine if there is evidence that

[1] For simplicity, the present study uses the language of PDSA cycles for all examples. The same arguments should apply to other similar iterative methods, e.g., the data wise improvement process.

the change idea led to improvement. Then, the evidence is used during the 'Act' phase of the cycle as a warrant to abandon or pursue the change idea.

The ability to quickly provide a strong warrant is integral to the success of this approach (Reed & Card, 2016). In industry, the primary quantitative method for providing this warrant is statistical process control (SPC) – a set of statistical diagnostic tests designed to be used by engineers in monitoring machine processes (Langley et al., 2009; Shewhart, 1931). Unfortunately, in education, processes are often complicated by human social dynamics (Koopmans, 2020; National Research Council, 2002) potentially making traditional SPC unworkable. Furthermore, many methods in SPC demand a statistical skill set that is less prevalent among educators (Chirume, 2018; Roderick, 2012).

Presently, the only suggested method for establishing quantitative evidence of improvement in education is the run chart (Bryk et al., 2015) – an enhanced point and line plot with statistical decision rules for estimating whether data is changing historically. Run charts are the least technical version of SPC. But they are still afflicted by the same core assumptions regarding data (e.g., exchangeability) which have not been closely examined in the educational context. Many data sources in education experience substantial autocorrelation (e.g., attendance) and nonstationarity (e.g., academic growth) (Koopmans, 2020) which could be problematic for the use of run charts. Moreover, the decision rules for run charts recommend collecting at least 10 data points and preferably 20-30 data points (Anhøj & Wentzel-Larsen, 2018; Langley et al., 2009). For sources of educational data collected weekly or slower, this could mean waiting three or more months to substantiate that a change was an improvement.

This is an untenable time scale for a method like the PDSA which is based on iteration, especially in education where improvement work is likely constrained by the school calendar.

The run chart is also only sparsely mentioned in the educational literature on continuous improvement, and, to my knowledge, the decision rules for run charts have never been fully explicated in the educational literature. While this literature clearly notes the critical importance of substantiating improvement with data and analytics, the same literature provides very few details on how this can or should be done. There may be a tacit assumption that the analytic tasks of continuous improvement can be met with bespoke solutions contrived by applied education researchers working either in or around schools.

Notwithstanding the potential logistical problems of this assumption, evidence from continuous improvement in health care already suggests this assumption may be flawed. In a systematic review of the use of the PDSA cycle across a decade of health care improvement work, only 15% of reviewed studies used quantitative data at monthly or more frequent intervals (Taylor et al., 2014) despite a rich literature in health care on methods for testing change ideas during PDSA cycles (Provost & Murray, 2011) including formal research designs (Speroff & O'Connor, 2004), control charts (Matthes et al., 2007), and run charts (Perla et al., 2011). Education is even more divorced from the industrial origins of continuous improvement than health care furthering the challenges of drawing methodological analogs. Moreover, applied education researchers are far less familiar with continuous improvement than health care researchers and, by comparison, there is little field specific literature on continuous improvement in education.

Given this it is perhaps unsurprising that improvement work in education thus far has not reported rigorously substantiating improvement during the PDSA cycle. While there are discussions of run charts or at least analyses in the formative texts of most continuous improvement methods (for an example from improvement science, see Bryk et al., 2015), these discussions tend to be limited to process measures which are proximal to the improvement work and more closely resemble implementation measures than outcome measures. In the formative texts and early literature on educational improvement projects there are no examples of run charts or rigorous PDSA analyses on important outcomes like teacher well-being, student attendance, or student literacy or mathematics growth.

While the lack of examples of run charts in concept and use is problematic in and of itself, this is further complicated by the potential for run charts to be inappropriate for use with some educational outcome measures. This is because many education measures are likely to have inherent dynamicity (e.g., trends, seasonality) which would require additional modeling beyond a simple run chart. Furthermore, even were run charts constructed appropriately for these measures, the current educational improvement literature provides no advice on how to determine if change has occurred. Without decision rules or diagnostic tests, run charts are simply a point and line plot with a median that must be idiosyncratically interpreted by the user. Under these conditions run charts are no longer a disciplined method for substantiating improvement.

Instead of rigorously evaluating changes during the PDSA cycle, educational improvement work is most often evaluated rigorously after the fact with quasi-experimental methods that are well understood by applied education researchers (for an example from

improvement science, see Yamada & Bryk, 2016). These evaluative methods provide evidence of the efficacy of continuous improvement as a reform strategy but are too lagged to substantiate improvement in the field (Bryk et al., 2015). This is not a criticism of early improvement work in education. Improvement workers certainly only pursued change ideas which were showing evidence of success. However, careful consistent methods for establishing and reporting evidence of improvement during a PDSA cycle are essential to rigorous, high-quality continuous improvement work (Reed & Card, 2016). Without disciplined methods, continuous improvement risks becoming another educational initiative where the results in practice vary greatly.

To realize the gains seen from continuous improvement in industry and health care, education must have disciplined methods for substantiating improvement during PDSA cycles that are appropriate for use in school contexts. To this end, the present study asks three initial questions which will help to inform future work in developing these methods:

1) What are the current methods used for substantiating improvement in industry and health care?
2) What are the differences between education and sectors like industry and health care that affect the feasibility of substantiating improvement using existing methods?
3) What characteristics are necessary in a method for substantiating improvement in education?

This study is organized as follows. In the first section, I present an overview of the literature on run charts and control charts, as these are the predominant methods for

substantiating improvement in industry and health care. In the second section, I examine how people, organizations, data, and research differ between education and other sectors like industry and health care. I also begin to consider the consequences of these differences for substantiating improvement. In the third section, I draw on the first two sections to theorize the methodological requirements for substantiating improvement in education. Finally, I conclude with a consideration of the deficiencies of available methods.

Methods for Substantiating Improvement in Industry and Health Care

> "When numbers are large, chance is the best warrant for certainty."
> Eddington (1929, p. 64)

The primary methods for substantiating improvement in industry and health care are run charts and control charts, also known as statistical process control (Langley et al., 2009; Provost & Murray, 2011). Walter Shewhart developed the concept of statistical process control while working at Bell Laboratories in the 1920s. Science, at the time, was beginning to recognize the importance of probability and statistics in understanding natural phenomena. Shewhart applied this burgeoning idea to the field of engineering, specifically quality control in manufacturing. Previous generations of quality engineers had envisioned a future where machines behaved exactly as directed producing identical products on each run. Shewhart argued that modern science did not support this vision. Moreover, the pursuit of perfect understanding was not even the goal of engineering. Shewhart proposed that quality control should embrace probability and statistics. A quality product, instead of having exact characteristics, could be one where the characteristics fall within some expected, reasonable level of variation

(Shewhart, 1931). These ideas, developed over the next decade, became the foundation for W. Edwards Deming's quality improvement work (Deming, 2000).

Interestingly, statistical process control was not created for the purpose of substantiating improvement, although, quality control is, ultimately, about improving quality. Shewhart was mostly interested in understanding and predicting quality. For Shewhart, the goal was to gain control of the process. When a manufacturing process was 'in control,' the quality of the product expected could be predicted using statistics. From there, expectations could be raised or narrowed if quality improvements were needed, or expectations could be held constant to watch for process degradation. Simply said, statistical process control was a method for detecting when a process changed. Whether the change was an improvement was not initially of consequence to Shewhart. In truth, the word 'improve' is mentioned less than a dozen times in Shewhart's formative publication on statistical process control (Shewhart, 1931).

At the same time Shewhart was developing statistical process control, the statisticians Jerzy Neyman and Ronald A. Fisher were popularizing the randomized controlled trial, mostly in their work in the field of agriculture (Fisher, 1937). This parallel development is important because the randomized controlled trial was another contender for substantiating improvement. Both techniques are experimental tests of change. The main difference between the two is the counterfactual approach. In a randomized controlled trial, a control group is used to estimate the counterfactual. Whereas in statistical process control, historical data of the process is used to estimate the counterfactual. This difference in estimating the counterfactual results in a key differentiation in purpose for the two methods.

Randomized controlled trials are useful for confirming the causal effects of treatment in a sample of a population. This made them ideal for agriculture and medicine where the potential for random sampling could be leveraged to generalize effects. Note, randomized controlled trials have struggled to replicate when performed with convenience samples, especially in fields like education where there is evidence of substantial effect heterogeneity (Weiss et al., 2017; for an example see Hanselman et al., 2017). In contrast to the randomized controlled trial, the purpose of statistical process control is to confirm the causal effect of treatment for a single subject (or potentially an aggregate of subjects). Generalizability is not a concern of statistical process control, as knowledge in quality control is built through frequent replication not statistical equivalence. Statistical process control was ideal for manufacturing where samples were often purposeful and small, e.g., a handful of machines producing products.

Today, run charts and control charts are the primary methods for substantiating improvement in industry and health care (Langley et al., 2009; Provost & Murray, 2011), and randomized controlled trials, particularly factorial designs, are occasionally a late chapter in a quality improvement text. In the sections that follow, I will provide a summary of the literature on control charts and run charts. The goal is to give the reader a better understanding of the strengths of these specific methods of statistical process control as well as their limitations.

Control Charts

A control chart is an enhanced point and line plot of a signal over time. In addition to the signal data, control charts include a center line and upper and lower control limits for the signal; typically, these are drawn as horizontal lines on the same plot. Traditionally, the center

line is the expected mean, and the upper and lower control limits are three standard deviations above and below the mean, respectively. To substantiate improvement, decision rules – tests which compare the signal data to the control limits – are applied to the control chart. For example, a common decision rule asks if any points in the signal fall outside of the control limits. If any of the decision rules test positive then there is evidence that the process is changing, and if the signal is moving in a preferable direction, then the change is considered an improvement.

The control chart makes two important assumptions. First, it assumes the data is ergodic meaning it does not exhibit dynamicity (Koopmans, 2015). This makes the data exchangeable. Second, the control chart assumes the distribution of the data is well-defined. Though, control charts have been shown to be robust to violations of this second assumption (Stoumbos & Reynolds, 2000). Given these assumptions, the counterfactual for the experimental data is assumed to be a probability distribution parameterized by estimates from some baseline period of observation. Decision rules, then, are essentially null hypothesis significance tests using the estimated counterfactual as the null condition. Deming would almost certainly object to this comparison as control charts are more heuristic than the comparison allows (Woodall, 2000). But I would argue it is a useful comparison which is only problematic when null hypothesis significance testing is incorrectly elevated above other diagnostic tests.

The most widespread control charts in use today are Shewhart control charts. Designed by Walter Shewhart in the 1920s, these control charts were the original tools of statistical process control and have not changed substantially in the past century (Stoumbos et al., 2000).

There are many different types of Shewhart control charts, typically chosen based on the distribution of the data being collected from the process. The most common chart in use is the X-bar chart which monitors the means of samples of a continuous variable, e.g., weight, temperature. Figure 2.1 presents an example of a Shewhart X-bar chart created by the R package qcc (Scrucca, 2004). Other Shewhart control charts are available for monitoring the standard deviations of samples (s chart), the time between rare events (T chart) as well as discrete variables including counts (C chart), rates (U chart), and proportions (P chart). There is likely a control chart available for almost any defined probability distribution (see Montgomery, 2007 for more types of control charts).

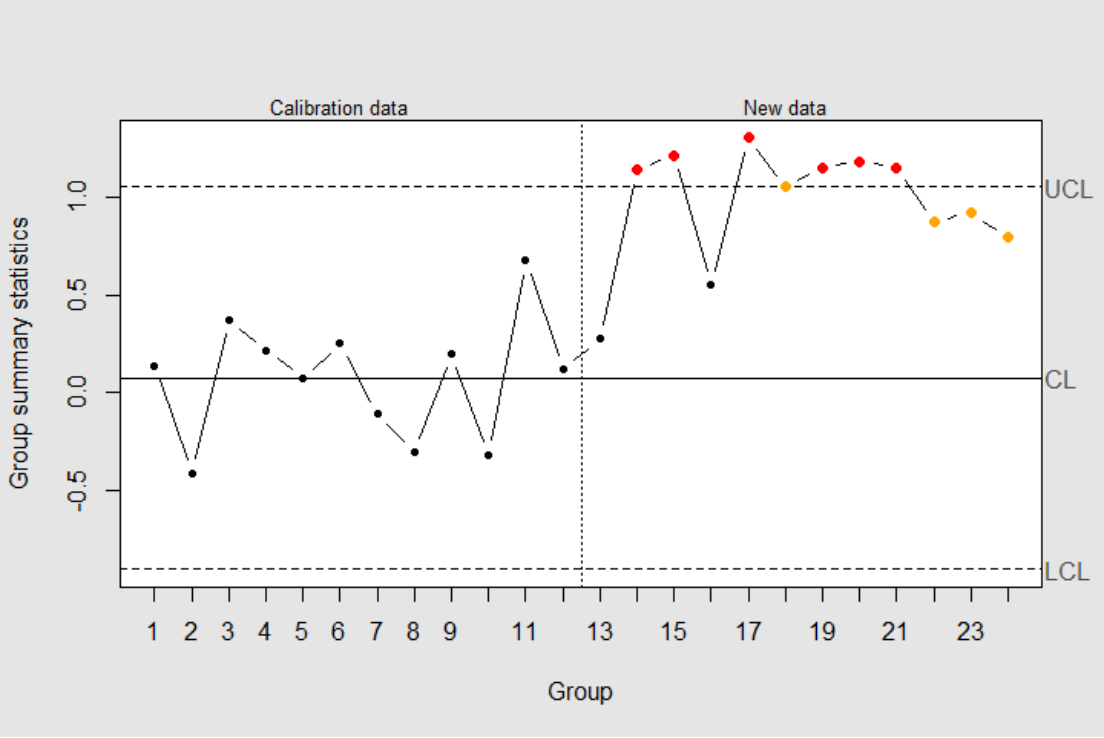

Figure 2.1. An example of a Shewhart X-bar chart where non-random change is present. The black line graph is the mean outcome data for the samples (N=10, M=0, SD=1). The dashed vertical line is the beginning of the non-random mean shift (D=1 SD). The solid horizontal line (CL) is the mean of the calibration data. The dashed horizontal lines (UCL & LCL) are the control limits for an X-bar chart (3 standard deviations). The red and orange dots in the new data are out of control signals.

The Shewhart control chart is best at detecting large shifts in the process mean (Stoumbos et al., 2000). Even brief spikes can usually be detected. However, a slow drift in the process mean can be difficult to see quickly using a Shewhart control chart. This is in part because the Shewhart control chart and its basic decision rules only use the current sample as a comparison. Consequently, evidence cannot accumulate over time (Stoumbos et al., 2000). Additional decision rules based on the statistics of runs (Z. Chen, 2010; Schilling, 2012) have been created to improve the performance of the Shewhart control chart (see Champ & Woodall, 1987). But modern statistical process control literature has largely moved away from studying Shewhart control charts, particularly for the task of detecting process drift and other gradual phenomena.

The CUSUM (Barnard, 1959; Page, 1954) and the Exponentially Weighted Moving Average (EWMA) chart (Roberts, 1959) are two equally performing and equally popular alternatives to the Shewhart control chart (Montgomery, 2007). These charts are direct replacements for the X-bar control chart and can also be adapted to other types of Shewhart control charts. Basically, the data is transformed to include some portion of the history in each point. For the CUSUM, the observations are normalized and then each sample is replaced by the sum of all samples up to that point in time. If the process has a constant mean, then these sums should hover around zero within some defined control limits. The EWMA chart, similarly, replaces each sample with an exponentially weighted moving average of all prior sample means. The CUSUM and the EWMA chart are best at detecting small persistent shifts in the process mean (Stoumbos et al., 2000) and, generally, perform poorly when detecting transient

shifts. Figure 2.2 presents an example of an EWMA chart created by the R package qcc (Scrucca, 2004).

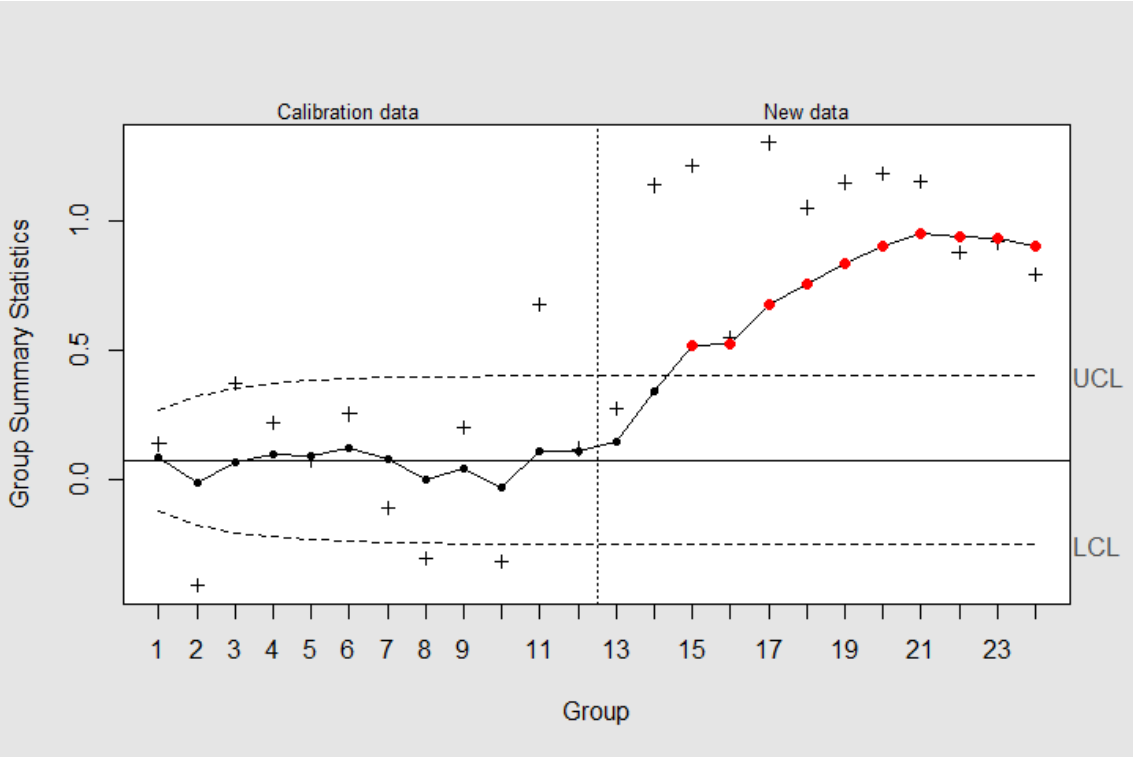

Figure 2.2. An example of a EWMA chart where non-random change is present. The black line graph is the exponentially weighted moving average of the mean outcome data for the samples (N=10, M=0, SD=1). The dashed vertical line is the beginning of the non-random mean shift (D=1 SD). The solid horizontal line (CL) is the mean of the calibration data. The dashed horizontal lines (UCL & LCL) are the control limits for an EWMA chart (3 standard deviations). The red dots in the new data are out of control signals.

Most of the research in statistical process control in the last few decades has focused on enhancing extant control charts to meet modern challenges (Woodall, 2000; Woodall & Montgomery, 2014). There have been two main thrusts to this work. The first is multivariate control charts. Often, in modern industrial applications, machines are monitored by hundreds of sensors. Instead of having hundreds of control charts, these methods allow engineers to

monitor multiple sensors as a composite measure with a single control chart (Ferrer, 2014). The second focus of recent statistical process control research is synergistic control. Although machine driven processes are often well defined and stable, machine calibrations drift, some physical and biological processes exhibit cycles or seasons, and human operators are often still involved. Consequently, even manufacturing data can exhibit dynamicity. Synergistic control is a method where time-series models are used to remove the dynamics from data before applying control charts to the model residuals (De Ketelaere et al., 2011).

It is important to note that quality improvement workers out in the field have been slow to adopt the innovations in control charts described above (Woodall, 2017, 2000). Synergistic control and multivariate control may feature prominently in the last few decades of literature, but they are not used broadly. Even CUSUM charts and EWMA charts still see relatively sparse use in practice given they have been available for over fifty years now (Stoumbos et al., 2000). Woodall (2000) attributes the stagnancy of statistical process control in the field to a number of factors including the cumbersome bureaucracy of quality assurance, the volume of experience quality improvement workers have with existing straightforward methods, the weak statistical backgrounds of quality improvement workers, and the disorganized and underdeveloped quality improvement literature base. In short, the Shewhart control chart, despite its weaknesses, has been the dominant method for substantiating improvement in industry for the last century.

Run Charts

The run chart, much like the control chart, is an enhanced point and line plot of a signal over time. However, the run chart uses the expected median for the center line, to reduce the influence of outlying observations, and does not include control limits. Instead, run charts rely on the statistics of runs – sequential points on one side of the median – to detect non-random variation (Z. Chen, 2010; Schilling, 2012). For example, the Anhøj decision rules for run charts asks two questions about runs (Anhøj, 2015):

1) Is the number of runs less than would be statistically expected given the quantity of data?
2) Is the longest run beyond what would be statistically expected given the quantity of data?

Similar to control charts, if any decision rules test positive then there is evidence that the process is changing, and if the signal is moving in a preferable direction, then the change is considered an improvement. The only assumption of the run chart is that the data is ergodic. Figure 2.3 presents an example of a run chart created by the R package qicharts2 (Anhøj, 2018).

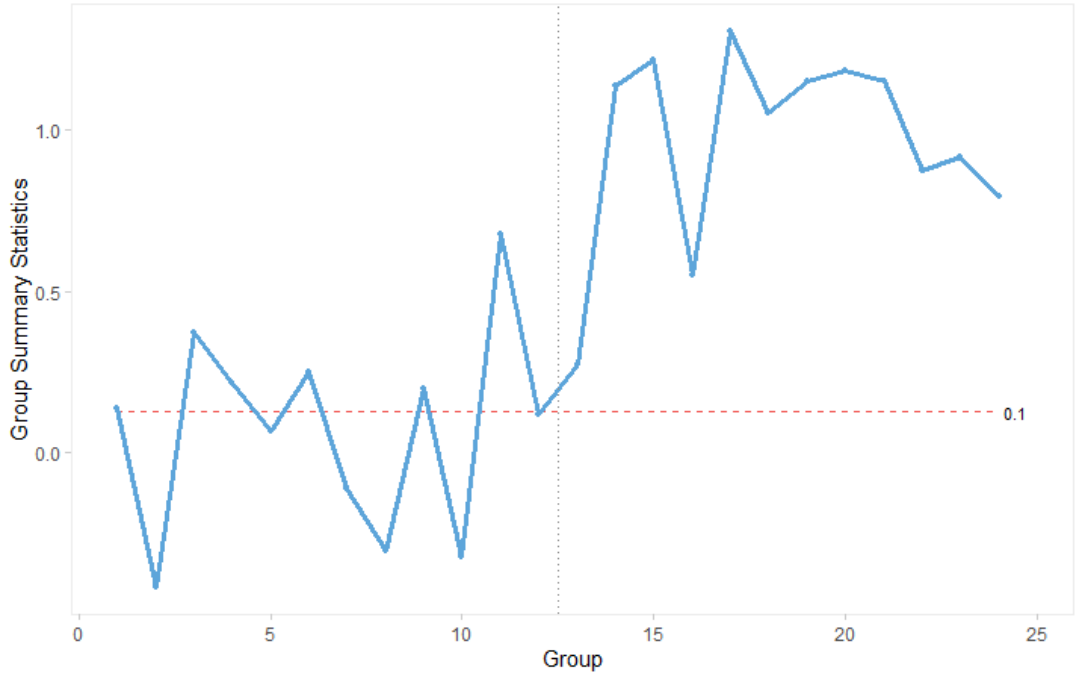

Figure 2.3. An example of a Run Chart where non-random change is present. The blue line graph is the raw outcome data (N=10, M=0, SD=1). The black dotted vertical line is the beginning of the non-random mean shift (D=1 SD). The red dashed horizontal line is the median of the calibration data.

The run chart is often considered a prototype for a control chart (Carey, 2002) or a practical alternative (Anjard, 1995), but has some clear advantages over control charts. First, the run chart is much easier to construct, interpret, and use than a control chart. The statistical portions of run chart decision rules can be looked up in a short table, and there is an abundance of clever techniques in the literature for smoothing the mathematical edges of run chart use. For instance, Carey (2002) points out that the median can be found simply by covering the run chart from the top down with the edge of a sheet of paper until half of the points are obscured. The second advantage of a run chart is it is less parametric than a control chart. Whereas there are many different types of control charts to address all manner of distributions of data, there is only one type of run chart, and the decision rules only assume that the distribution is ergodic

and somewhat symmetrical. Anhøj & Wentzel-Larsen (2018) suggest the run chart is best used as a simple yet robust initial test of change.

Run charts are less prominent in the statistical process control literature from industry, but in health care, improvement scholars have been advocating for the use of run charts for years (Perla et al., 2011). Unfortunately, as was discussed in the introduction, Taylor et al. (2014) found only 15% of studies in a decade of health care improvement work collected data frequently enough to justify a run chart. In practice, the run chart may be used more often than has been estimated. But it is noteworthy that there is a parallel between health care's struggles to adopt basic improvement methods and industry's lethargy in progressing theirs.

As a general principal, methods are designed with assumptions based on a specific context. Shewhart's control charts have been used for a century in industry because they worked very well in the manufacturing context where they were introduced. From Deming's perspective there was no better method.

> "The Shewhart control charts do a good job under a wide range of conditions. No one has yet wrought improvement." (Deming, 2000, p. 180)

Some scholars have seen Deming's view as a contradiction, given its conservatism on improvement (Woodall, 2000). I would suggest, instead, that it is an outlook born of pragmatism in the spirit of John Dewey (Dewey, 1910) which is unsurprising given Dewey's writings are an intellectual forebearer of much of what is called continuous improvement in education today (Yurkofsky et al., 2020). Deming experienced decades of success using Shewhart control charts in manufacturing. He was understandably hesitant to suggest an

alternative may be needed given his real-world experiences. The question is whether Deming's experiences translate to new contexts in modern industrial applications, health care, and education. In the next section, I will explicate the differences between education and other sectors which may challenge the assumptions of traditional statistical process control.

Differences between Education and Other Sectors

Education is a sector with many different aims and ends. This is in part because educational systems serve multiple functions in democratic societies (G. Biesta, 2009). In a framework for discussing the purpose of education, G. Biesta (2009) identifies three broad functions for education. First, education serves a qualification function by providing skills, knowledge, and other forms of learning. Second, education serves a socialization function whereby individuals become well-mannered members of society. Third, education serves an individuation function that acts opposite socialization with the goal of developing independent thinking. These functions are often at odds with one another which can lead to complex and competing goals in education, especially as the educational system interacts with a diverse group of stakeholders (National Research Council, 2002).

The goals of industry and health care are more nuanced than simply increasing profits and saving human lives. However, businesses tend to have goals that all lie under the umbrella of raising profits, and in health care there is an expectation that human health is the overarching concern. Education does not have this type of comprehensive focus. In recent years, there has been some alignment in direction due to accountability policies and measures

(G. Biesta, 2009). Valid criticism, though, has also been leveled at accountability which often undermines the discussion around the many varied purposes of education (G. Biesta, 2009).

Continuous improvement is attractive to educators because the methods attend to the complexity that arises from education's ambiguity, variability, and interdependence (Yurkofsky et al., 2020). In industry and health care, continuous improvement tackled thorny seemingly intractable challenges, e.g., improving physician hand hygiene (see White et al., 2012). But it did so in sectors where overall there was more cohesion. Moreover, quality improvement in industry and health care often focused on problems which could be compartmentalized. Education's multiplicity of purposes, heterogeneity of actors, and wholly interdependent systems, make it difficult to achieve the same conditions, especially at scale.

In the subsections that follow, I will examine the specific differences between education, industry, and health care. I will organize the differences under four categories: 1) the subjects of improvement, 2) data and measurement, 3) practitioners and the organization, and 4) research and development. As these differences are explored, I will also begin to consider the consequences of these differences for substantiating improvement.

Subjects of Improvement

In manufacturing the subject of improvement is typically some widget under production. Improving the quality of the widget involves controlling how many widgets fall outside of the specifications. However, widgets are relatively homogeneous even if they vary on some important physical dimensions during production. Also, widgets, generally, can be understood with the physical sciences and don't have the capacity to act independently.

Furthermore, a change which results in a bad batch of widgets only increases expenses because widgets only have monetary value.

Compare this to education where the subject of improvement is often the student. There are no well-defined specifications for students. Benchmarks around report cards, test scores, and attendance can serve some purposes here, but as was stated above, there are often competing and conflicting goals in education which complicate defining clear and shared standards. Students are also not homogeneous. They vary greatly even within schools (Rodriguez & Nickodem, 2018). Furthermore, students are people which, although governed by the physical and biological sciences, are best understood with the social sciences. And, importantly, a change which results in a 'bad batch' of students is never acceptable. Although there are many targets for improvement in education that are not students (e.g., teachers and administrators, educational routines and processes) any initiative in education must improve students at least indirectly.

In contrast to industry, health care provides for a better comparison. The subjects in health care are patients, who are people much like students. Patients' lives have more than monetary value, and patients are best understood with more than just the physical and biological sciences. However, patients do generally have well-defined specifications, especially as specialization in medicine limits the scope of concern for most health care providers. Also, although there are some variations in diagnoses, treatments, and care across people, humans have many physiological commonalities. Furthermore, a person's health can be quantified using primarily biological measures. The social sciences only are required for specific problems, e.g., the behavioral components of asthma care (see Bravata et al., 2009).

Industry, health care, and education form a continuum of sorts. For industrial applications, the subjects of improvement are well-defined, relatively homogeneous, and individually immaterial. On the opposite end, for educational applications, the subjects are ill-defined, exceedingly heterogeneous, and individually invaluable. This difference in the subjects of improvement has at least two important implications for substantiating improvement in education. First, since the subjects in education are children not widgets, experimentation during a PDSA cycle will face additional practical and ethical challenges. Second, since the subjects in education exhibit substantial variation and the targets for these subjects are ill-defined, common problems and solutions may be limited to small groupings, and scaling may require considerable adaptation. In sum, improvement in education must contend with the challenge of working with small purposeful samples of vulnerable human subjects.

Data and Measurement

Measurement in industry and health care is more straightforward than in education. This is in part because of the ambiguity in education's goals. However, it's also largely due the proximity of industry and health care to the natural sciences. Industrial applications can track physical characteristics, e.g., the weight of a widget, and health care applications can track biological characteristics, e.g., the blood pressure of a patient. Education, even when it can agree on a goal, must contend with data that is generated by human volition (National Research Council, 2002). Consequently, educational measures often have substantial unexplained variation which must be modeled stochastically (Wooldridge, 2015), and the major causes of variation in educational data are often out of the control of the educational system (Berliner, 2013).

Educational measurement is also on a much slower cadence than measurement in industry and health care. This is mostly because physical and biological processes are more responsive to change. The responsiveness of a process is a product of the frequency of process cycles and the momentum of the process. Making a batch of widgets can occur in minutes. Attending school can only be observed daily, and learning occurs best over weeks or even months (Cepeda et al., 2006). Moreover, each batch of widgets has no memory of the last batch. Education, though, is replete with feedback loops (Koopmans, 2020) which can retard the pace of change further. All of this is then compounded by substantial variation in educational data which decreases the resolution of educational measures. As a result, useful educational data is created and collected much less frequently than data in industry or health care.

Another important consideration when measuring people, as opposed to machines or human bodies, is the propensity for change over time. Dynamicity is a growing concern in data from industry and health care (Box & Narasimhan, 2010). But, typically, in these sectors dynamics are complications not expectations. In education, students are expected to grow over time as they learn, each school year always begins a new cycle of behaviors, and disruptions and irregularities are the norm. Consequently, many data sources in education exhibit non-ergodic variation including trends (e.g., academic growth), seasons (e.g., absenteeism), and outlying observations (Koopmans, 2015; Koopmans, 2020). Together these differences suggest improvement in education should anticipate data which features unexplained variation and complex temporal dynamics as well as infrequent and often limited collection.

Practitioners and the Organization

Practitioners in education are different from industrial engineers and health care providers. Preservice training for education is professional in nature, much like engineering and medicine. However, the body of knowledge conveyed to educators is not as uniform as it is in other professions (Cochran-Smith, 2005; Labaree, 1992). For instance, some educators will hear about continuous improvement in school, but many will not. One important area where educators will vary greatly is in their capacity for statistics and research. Math and science teachers may have been educated in this area, and an estimated quarter of teacher preparation programs offer coursework in this area (Chirume, 2018). However, overall, educators are likely to have much less training and inclination to work with statistics and research than industrial engineers and health care providers (Roderick, 2012).

Educators are also different in that they are not fully professionalized (Ingersoll & Perda, 2008). In the literature on professionalization, education is referred to as a semiprofession (Lortie, 1975). This is in part because teachers are held accountable by administrators instead of the professional body. However, educators also have incredible autonomy in decisions regarding their classrooms and students due to the loosely coupled organizational structure of schools (Weick, 1976). Educators often leverage this distinction to ignore orders from their superiors by complying with the letter of the order but not the spirit (Meyer & Rowan, 1977). This is of consequence for improvement work because there is the potential that educators will ignore improvement work if they find it to be cumbersome or useless. And, importantly, there is limited professional accountability that can accelerate their acceptance of continuous improvement in education.

Educational organizations are also different from organizations in industry and health care. Most obviously, the majority of schools are under public governance. Thus, there is a political dimension to improvement work in education (Yurkofsky et al., 2020). An educator's time is always in high demand from many fragmentary layers of the educational system (Chubb & Moe, 1991). New initiatives in education must quickly justify their use of resources to those at the bottom and the top of the bureaucratic hierarchy. For practitioners, as was suggested above, initiatives need to be pragmatic. But, for educational policy makers and leaders, initiatives need interpretable, actionable evidence (Coburn et al., 2009). And, after years of accountability pressure (Mehta, 2015), the standards of evidence in education are increasingly rigorous (Slavin, 2020).

In summary, improvement in education should be prepared to work with educators who are not trained to meet the analytic needs of substantiating improvement and also have little time to spare for improvement work unless it is relevant to their daily practice. Improvement in education should also be prepared to convince leaders of its value. Strong evidence during improvement work will be important in making this case.

Research and Development

Educational research and development is not funded to the same degree as research and development in industry and health care (Bryk & Gomez, 2008). In 2018, total U.S. medical and health research and development spending was $194.2 billion which is about 5% of total U.S. medical and health spending (Research!America, 2019). In education, research and development accounted for only $1.5 billion which is less than 1% of total U.S. educational

spending (Gibbons, 2020). Moreover, much of this funding supports theory development not practical research (Bryk & Gomez, 2008). Consequently, education relies on rigorous standards of evidence to evaluate the glut of potentially underresearched and underdeveloped educational solutions available on the market (Slavin, 2002). As was noted above, the scarcity of an educator's time and accountability politics have played a role in escalating standards of evidence in education. However, these factors are, in truth, manifestations of the limited resources in education which are most stark in the area of research and development.

Educational research also differs from research in industry and health care in the expectations for the size of average treatment effects. Control limits in industry are typically set to three standard deviations. This might seem outrageous to educational researchers, who often power their studies to find effects which are 30 times smaller. But, in industry and health care documented effects are large compared to education. In prior sections, this issue has already been raised to a degree. Education is a field characterized by human volition, exceptional variation, and a diversity of actors (National Research Council, 2002). Consequently, education is difficult to control. However, I contend that the variability in education is more of a problem for large random samples and cross-sectional analyses. To understand why consider the example of one-on-one tutoring. Why is it far more effective than the traditional classroom experience (Chi et al., 2001)? I would suggest that the tutor has better control over the student's learning experience resulting a more student-centered approach. Single subject designs, e.g., time-series analyses, from the subfield of special education have leveraged this difference for decades to conduct high-quality research on individualized educational needs. Unsurprisingly, single subject educational experiments often see effects larger than two

standard deviations in size even in meta-analyses (e.g., DuPaul et al., 2012; Gierut et al., 2015; Lee Swanson & Sachse-Lee, 2000).

Educational improvement work will need to experience success using rigorous methods to justify its existence. Improvement workers, though, should not be too concerned with the historically small effects in educational research. When considering the potential for change it is folly to limit the possibilities based on what happened on average in the past (Payne & Ortiz, 2017). Especially, given the areas of education, like special education, that look most like improvement work have seen different results.

This subsection and those above have summarized some of the key differences between education and other sectors that may be of consequence for substantiating improvement in education. This explication has not been exhaustive in nature. But, instead, aimed to show the contours of the issues that might arise from directly transporting statistical process control into educational improvement work. In the next section, I will develop methodological requirements for substantiating improvement in education that address these issues.

Methodological Requirements for Substantiating Improvement in Education

The difficulty in specifying methodological requirements for substantiating improvement arises foremost from the generality of the task. Research methods are chosen based on the research question (National Research Council, 2002). Thus, assuming a common form of question in continuous improvement is the first step in motivating the methodological requirements of improvement.

Scientific questions are formulated in terms of a dependent variable, an independent variable, and a subject. In a PDSA cycle, after a change idea has been enacted, evidence is gathered and analyzed to determine if the change was an improvement. This statement can be rephrased as the following question:

> Did the change idea *(independent variable)* improve the outcome *(dependent variable)* for the people targeted *(subjects)*?

This is a testable scientific question under certain assumptions, namely:

- The **people targeted** are well specified. *e.g., students at Bayside High School who were chronically absent in the first semester and also have Asthma*
- The **change idea** is well defined and motivated by theory (Lipsey, 1993). *e.g., new protocol for the school nurse in coordinating care for students with Asthma designed to reduce Asthma related absence*
- The expected result of the change is valued and specific, i.e., an **improvement**. *e.g., a decrease in absenteeism for students previously identified as chronically absent and Asthmatic*
- The **outcome** will contain a measurable signal of the expected result of the change. *e.g., the attendance of chronically absent students who have Asthma*

Researchers regularly make similar assumptions to justify their work (Lipsey, 1993). However, there are a few notable differences between improvement work and academic research. First, in improvement work the intended result of the change idea is motivated by values. This is a key distinction with academic research where the quest for understanding

often trumps values. Second, the change idea is based on a working theory of practice improvement (Bryk et al., 2015) which often includes academic theory but is specific to the context where the change will occur. The key distinction here is the scope of the knowledge gained. Improvement work is not immediately concerned with generalizing beyond the people targeted. Instead, generalizability is expected to slowly build over time by spreading change ideas to new contexts and adjusting them in response to failure (Bryk et al., 2015). As improvement work aims to both improve valued outcomes and build knowledge – namely a working theory of improvement – it necessarily straddles the line between pure applied research and pure basic research. This class of research is often referred to as Pasteur's Quadrant (Stokes, 2011).

For an improvement worker who is conducting a PDSA cycle, the above research question must be answered to move forward. The iterative nature of the work requires that a decision be made, based on evidence, whether to continue, change or stop the improvement cycle. In industry and health care, statistical process control is the quantitative method used to address this question. A control chart or a run chart is built using time-series outcome data from the subjects targeted. If the chart's decision rules indicate that nonrandom variation was present after the change idea was introduced and the outcome is moving in a valued direction, then there is evidence the change was an improvement.

To answer the above question in educational improvement work an appropriate method is needed. Control charts and run charts can work in some situations in education. But, as was shown above, education has key differences in people, organizations, data, and research which may challenge the assumptions of control charts and run charts. Even so, statistical process

control is still the best starting point for a method to substantiate improvement in education for at least three reasons: 1) It works well with small purposeful samples. 2) It is a causal method when used in experimental contexts like the PDSA cycle. 3) Control charts and run charts are relatively practical for field use.

Considering the important features of statistical process control and the differences between education and other sectors, I identified five methodological needs for substantiating improvement in education. These needs are as follows:

The method must be…

1) Disciplined but pragmatic
2) Appropriate for use with small samples of subjects
3) Responsive with limited time-series data
4) Semiparametric
5) Unobtrusive and automated

Some of these needs are already attended to by statistical process control, and education merely has the same or a greater need than industry or health care. However, others of these needs are in direct conflict with the assumptions of control charts and run charts. In the sections that follow, I will further motivate each need and consider how statistical process control may or may not address the need.

Disciplined but pragmatic

The first requirement of a method for establishing evidence of improvement in education is that it must be disciplined but pragmatic. This requirement has two equally important halves. First, the method must be scientifically rigorous. The simplicity of the PDSA cycle belies the difficulty in successfully applying it (Reed & Card, 2016). The PDSA cycle and continuous improvement are demanding scientific methods (Bryk et al., 2015; Deming, 2000; Langley et al., 2009; Shewhart, 1931) and as such are predicated on establishing causal relationships (Dewey, 1910). The challenges of substantiating improvement are the same challenges faced in applying the scientific method. Consequently, a scientifically rigorous method with appropriate assumptions is essential to substantiate improvement in education.

Establishing evidence using a disciplined method benefits continuous improvement in education in a number of ways. First, it guards against bias. As Dewey (1910) wrote in *How We Think*:

> All the instrumentalities of observation … fill a part of their scientific role in helping to eliminate meanings supplied because of habit, prejudice, the strong momentary preoccupation of excitement and anticipation, and by the vogue of existing theories. (p. 80)

Education is replete with entrenched practices and ideas which persist through a multitude of isomorphic processes (DiMaggio & Powell, 1983). Furthermore, educators are overworked and often burnt out which can lead to cynicism (e.g., we tried that before, and it won't work) (Salanova et al., 2005). A causal method can help ward off false beliefs that persist across the institution of education.

A disciplined method is also important in education for political reasons. As was noted prior, education is largely under public governance, and education's limited resources (Bryk & Gomez, 2008) together with a history of increasingly top-down accountability (Mehta, 2015) has led to a challenging marketplace for new ideas. Presently, ESSA mandates that education programs be evidence-based – meaning there is extant evidence that the program has been successful. The implication of this mandate is that scientific rigor is vitally important for sustaining improvement work in the current political climate in education. In fact, in the Department of Education's non-regulatory guidance for states' implementations of ESSA, the highest level of effectiveness an evaluation can award – strong evidence – is only merited in pure experimental studies. Methodologies for substantiating improvement should aim to be as scientifically rigorous as possible.

Deming (2000) described continuous improvement as a system of profound knowledge that generates theory which can be used to understand organizations. Disciplined methods are fundamentally valuable to continuous improvement because improvement work is theory building, and to build scientific theory, even hyperlocal working theory, requires attention to the scientific method. Just as in the formal sciences, no single test can ever confirm a hypothesis (Dewey, 1910). However, many PDSA cycles can together result in an accumulation of evidence that can be used for rational prediction (Deming, 2000).

The second half of this requirement states that the method must also be pragmatic. This is an important but unremarked strength of statistical process control methods like control charts and run charts. Often run charts are described as practical because they have less overt statistical methods. However, both control charts and run charts possess other qualities which

make them practical for field use. For example, the incremental creation of data visualizations in these methods invites the user to eagerly engage with the data as the chart is constructed, and the simple decision rule based diagnostic tests quantify the user's observations quickly encouraging prompt and thoughtful reflection. Pragmatism is even more important in education than other sectors because practitioners in education can close their doors if an initiative feels impractical (Meyer & Rowan, 1977). To avoid a future where continuous improvement is merely ceremonial (Yurkofsky et al., 2020), improvement methods must empower educational practitioners by giving them rapid, useful feedback.

Appropriate for Small Samples of Subjects

The second requirement of a method for establishing evidence of improvement is that it must be appropriate for use with small samples, ideally single subjects. This requirement is not unique to education. Statistical process control was created to address similar needs in manufacturing. However, this requirement may be even more pressing in educational contexts.

As was noted prior, students are not like widgets on an assembly line or even patients in hospitals. They are exceedingly heterogeneous. Variability and diversity are well documented fixtures of education (National Research Council, 2002). Given the heterogeneity in students, the problems and solutions in continuous improvement will often be localized. For example, a student might not make it to school because their little brother was sick and they had to stay home to take care of him, or a student may skip school because they didn't want to take a math test they weren't prepared for. As problems become more localized the number of subjects the solution is appropriate for decreases. Even though there might be hundreds of students who

are chronically absent, there might be only a dozen students whose attendance is improved by a specific change idea. Over time, applying many small solutions will accumulate into a big improvement, especially when these changes are high leverage (Bryk et al., 2015). There is evidence that the dynamics of schools can be recursive and quick wins at key times can change the trajectory of students' school experiences (Yeager & Walton, 2011; for an example, see Borman et al., 2018).

The scope of continuous improvement further reinforces the need for small samples. As was discussed above, this is a key difference between scientific research and continuous improvement. The goal in continuous improvement is to learn quick about a system by making rapid changes that, optimistically, succeed, but often fail (Bryk et al., 2015). When changes don't work as expected, knowledge of the system is gained. Basically, the quicker improvement workers fail the quicker they learn. In improvement, changes are kept small in scope, size, and duration to increase the speed of change cycles.

Small changes are also less harmful when they do not turn out as expected. Education has consequences for failure. Students are vulnerable human subjects (United States, 1978) not widgets on an assembly line, and continuous improvement methods must protect the interests of students. Continuous improvement has great potential to benefit students, however, it could, as with all interventions, potentially harm them at times. To mitigate the impact of failure on students, change ideas should initially affect as few students as possible and for as brief a time as possible until evidence accumulates that the benefits of spreading the change are greater than the potential for harm.

Statistical process control addressed the limitations of small purposeful samples in manufacturing by collecting many observations of the samples over time. Educational improvement will likely need to take the same approach. The power of cross-sectional statistical methods is dependent on sample size, and inferential statistics are only consistent asymptotically (Wooldridge, 2015). Moreover, clustering, which is a defining characteristic of cross-sectional data in education (Raudenbush & Bryk, 2002), can raise additional problems in small samples (Imbens & Kolesár, 2016; McNeish & Stapleton, 2016). Therefore, the need for a method which is appropriate for small samples also likely necessitates the need for the collection of longitudinal data. Time-series of aggregate outcome statistics or panels of subject outcomes can be used to draw causal conclusions by leveraging exogenous variation over time (Morgan & Winship, 2015).

Responsive with Limited Time-series Data

The third requirement of a method for establishing evidence of improvement is that it must be responsive with only limited time-series data. As was identified in the discussion of the differences between education and other sectors, educational data is created and collected on a much slower cadence than data in industry and health care. This is because educational processes cycle less frequently (e.g., attendance only occurs daily), have more inertia (e.g., learning processes are constructivist), and have less resolution (i.e., small effects compared to variance) than processes in other sectors. As a result, educational measures are collected infrequently. Furthermore, although obvious but completely under remarked, the school year in education is relatively short, at least in the United States, and discontinuous compounding the problem. For example, a measure that is captured on a monthly cadence, will only have

nine data points available in a single school year; for a weekly measure only 36 data points will be available.

There is some potential for utilizing daily measurement in education. However, education does not have the measurement habits or practices to support this high a frequency of data collection. Health care was able to leverage an existing system of charting and documentation to support improvement work. There is no comparable system in education currently, and frequent data collection would need to be justified across an already resource strapped educational bureaucracy (Yurkofsky et al., 2020).

This is the first of the methodological needs which is unique to education. It is also a problematic requirement because improvement work is already conducted on small purposeful samples. Statistical process control, in industry and health care, builds statistical confidence by collecting ample historical data. Guidance for run charts and control charts, for instance, suggests collecting 20 to 30 data points (Anhøj & Wentzel-Larsen, 2018; Langley et al., 2009). A method for establishing evidence of improvement in education may not have this luxury.

Unfortunately, there are only a limited number of ways to increase statistical power. Generally, collecting more data is the first avenue. When this is not a possibility, power can also be increased by reducing the variance, increasing the effect, or accepting more false positives (J. Cohen, 1992). The PDSA cycle already leverages some of these avenues. Targeting and tailoring interventions can reduce variance and increase effects in the social sciences (G. L. Cohen et al., 2017). This should benefit the PDSA cycle which encourages small, focused change ideas which grow in scope as they adapt to new contexts (Bryk et al., 2015). Also, due to the

amount of replication in continuous improvement work PDSA cycles can likely accept more false positives than research, especially during the early stages of investigation. But these avenues on their own are unlikely to be sufficient to make up for the depleted time-series data that is most likely to be available in educational contexts.

Another avenue that should be considered is fully utilizing the limited data that is collected. Statistical process control was birthed well before the modern computer. Consequentially, there may be ways to leverage sophisticated statistical techniques to increase power, at least compared to traditional, archaic methods like the Shewhart control chart. For example, a weak Bayesian prior might raise the confidence of the counterfactual to acceptable levels despite only having a couple data points, or a forecasting technique might be used in place of decision rules to make a more powerful diagnostic test (Brodersen et al., 2015). This study does not have the answer for this challenge. But, given the data simplifications in decision rules for control charts and run charts, it seems likely there is more information in the data than is currently being used by traditional statistical process control.

Semiparametric

The fourth requirement of a method for establishing evidence of improvement is that it must be semiparametric. A semiparametric method is both parametric and non-parametric. This means part of the model is prespecified and part arises from the data. Linear regression, i.e., ordinary least squares regression, is an example of a classic parametric regression method where the data is fully parameterized by the regression coefficients and the variance of the residual error assuming normally distributed exchangeable residuals. A comparable

nonparametric regression method is local regression, e.g., LOESS (locally estimated scatter plot smoothing). LOESS can be thought of as many different piecewise linear regressions which together fit the data. This technique is non-parametric because the data determines the model. Assuming there is infinite data there would be infinite parameters.

Parameterized models are always an approximation of the truth (Murnane & Willett, 2010). When the mechanism being modeled is well understood a fully parametric model can be very predictive. However, underdeveloped, or incorrect model specification can bias prediction. Considering that educational improvement work's small samples will necessitate time-series analyses and the functional forms for educational time-series are largely unexplored, nonparametric methods are likely crucial to substantiating improvement using time-series data in education, especially if causality is important (Brodersen et al., 2015). Moreover, educational processes may not even have determinant functional forms. Instead, as was alluded to prior, daily attendance, disciplinary incidence rates, and many outcomes of teaching and learning likely exhibit complex temporal dynamics (Koopmans, 2020).

Unfortunately, given the substantial variation in data from educational contexts (National Research Council, 2002), a completely nonparametric model is likely also untenable. Evidence of improvement can only be established if the signal can be discerned from the noise. Normally an effect must simply be large in comparison to the variance of the data for it to be statistically significant. When time-series data includes dynamics, the effect must also become large quick enough in comparison to the dynamics of the data. A semiparametric approach is a good compromise. Adding in explanatory factors or control signals can increase statistical

power without bias, as long as residual variation and dynamicity is still appropriately modeled nonparametrically (Brodersen et al., 2015).

Control charts and run charts do not meet this requirement. The control chart is a fully parametric method where the chart type, center line, and control limits define the data's distribution, and the run chart is a largely nonparametric method because it works independent of the distribution of the data. The problem is control charts and run charts are both likely to be statistically weak in education. Control charts will be weak due to model misspecifications and run charts will be weak due to unaccounted variance. Furthermore, both control charts and run charts assume the data is ergodic. This assumption will often be wrong. Educational data features complex temporal dynamics (Koopmans, 2020). A semiparametric method for establishing evidence of improvement in education will be needed to mitigate these complications.

Unobtrusive and Automated

The fifth and final requirement of a methodology for establishing evidence of improvement is that it must be both unobtrusive and automated. An unobtrusive method for substantiating improvement does not ask more of educators than a PDSA cycle would, and an automated method is one that does not require any special statistical training. Research design elements like randomized treatments are too disruptive, and research methods which require an advanced degree to interpret the results are too restrictive. Basically, improvement methods need to be frictionless for educators. Ideally, a method for substantiating improvement would

be no more complicated than a medical diagnostic like a pregnancy test. Improvement work is enough of a challenge without methodological complications.

This need is distinct to education. As was discussed prior, educators have less training and inclination to work with statistics and research than industrial engineers and health care providers (Roderick, 2012). The data-driven decision-making movement in education has published extensively on education's struggles in this area (see Means et al., 2011). Also, engineers and doctors are trained in mathematics and the natural sciences. Many educators come from the humanities and social sciences. Furthermore, teacher preparation programs only rarely offer courses in research and statistics (Chirume, 2018). This is not to say that teachers are incapable. Teachers have many strengths which define their profession. However, they likely have less training and inclination to work with statistics and research than professionals in industry and health care. This need is also distinct to education because educators have the capability to disregard or undermine improvement work which they find burdensome (Meyer & Rowan, 1977). More tightly coupled organizational structures and strong professionalization decrease the ability of engineers and doctors to outright ignore cumbersome tasks. But educators are street-level bureaucrats with limited professional accountability who are accustomed to making their own decisions (Lipsky, 2010).

Unobtrusive automated methods are also important because the subjects of educational improvement are humans. Change ideas are mini-experiments and, consequently, will suffer from all the major post-treatment problems of experimentation (e.g., placebo effect, Hawthorne effect, John Henry effect, see Deaton & Cartwright, 2018). Methods for establishing evidence of improvement should not exacerbate the circumstances by requiring overt

alterations to normal school routines. Experiments work best when the subjects are blind to the experiment. The PDSA cycle in the context of a professional educators' normal work is well positioned to remain stealthy, so long as the method does not require otherwise.

Discussion

The present study aimed to answer three questions which would support future work to develop a method for substantiating improvement in education. First, this study asked:

> What are the current methods used for substantiating improvement in industry and health care?

I found that the most practiced method in industry was the century old Shewhart control chart from statistical process control. However, the quality improvement literature also included many modern alternatives which despite their strengths enjoy only niche usage. In health care, I found that the run chart was proffered as a practical alternative to the control chart, and perhaps saw more use. Though, there was also evidence in the health care literature that more traditional research methods are often employed.

Second, this study asked:

> What are the differences between education and sectors like industry and health care that affect the feasibility of substantiating improvement using existing methods?

I analyzed the differences in four categories, namely people, organizations, data, and research. I found impactful differences across all categories. Students, as the subjects of improvement, presented unique challenges compared to widgets and patients. Educational practitioners and

the educational organization were fundamentally different from practitioners and organizations in the more professionalized and cohesive industrial and health care sectors. Data in education was more complex and less plentiful compared to data in industry and health care. And research and development in education was poorly funded compared to research and development in other sectors.

Third, this study asked:

What characteristics are necessary in a method for substantiating improvement in education?

Based on the needs of statistical process control and the differences between education and other sectors, I identified five requirements of a method for substantiating improvement in education. 1) The method must be disciplined but pragmatic. The word 'improve' is causal language that can be translated as 'have a valued impact on.' 2) The method must be appropriate for use with small samples of subjects. The variation and diversity in education necessitates small purposeful changes. 3) The method must be responsive with limited time-series data. Educational processes accrete data on a much slower tempo than biological and industrial ones. 4) The method must be semiparametric. Education will need flexible methods to accommodate human volition and exceptional diversity. 5) The method must be unobtrusive and automated. Methods for establishing evidence of improvement should be seamless for educators, otherwise traction may be difficult.

The difficulty with this set of methodological needs is how contradictory they are. A disciplined method and a practical method are often making the opposite concessions. Limited

sample sizes with limited observations of each sample provide little room for statistical inference. Semiparametric methods are attempting to apply a model and elicit a model simultaneously. And all these divergent needs must be met without imposing on the educator. Consequently, most methods fall woefully short. Table 2.1 presents a comparison between the methodological needs for substantiating improvement and some common methods from educational research and statistical process control. The problem is the methods from educational research are not practical and too often rely on large cross-sectional samples, and the methods from statistical process control require too much time-series data and cannot accommodate education's considerable variation and dynamicity.

Table 2.1. Comparison Between Methodological Needs for CI in Education and Various Methods from Educational Research and Statistical Process Control

	Educational Research Methods			Statistical Process Control	
	RCT	Quasi-experiment	Single subject	Run Chart	Control Chart
Disciplined	X	X	X	X	X
Pragmatic			X	X	X
Small sample			X	X	X
Limited observations	X	X			
Semiparametric	X				
Unobtrusive		X		X	X
Automated				X	

Future research should consider two avenues for addressing these methodological needs. First, there could be methodological developments from applied math which have the

potential to meet these needs. Computational power is always increasing and exposing new discrete mathematical methods which previously would have required too much energy or time to consider. Second, there could be methods created for the technological sector or the financial sector which have the potential to meet these needs. One of the greatest challenges of addressing these requirements will be assembling an automated method that is semiparametric. Most complex models are the particular creations of talented scientists. But this challenge is likely already salient in fields where big data has been mainstream for decades.

To realize the gains seen from continuous improvement in industry and health care, education must have disciplined methods for substantiating improvement during PDSA cycles that are appropriate for use in school contexts. The requirements suggested by this study are challenging, but necessary. Improvement science is poised to become a major player in educational reform. Whether it succeeds at making meaningful, consistent improvement across the educational system is dependent on meeting this challenge.

Chapter 3 – Statistical Process Control using Bayesian Structural Time-Series Models

Introduction

Continuous improvement – a strategy for advancing organizational quality (Deming, 2000; Langley et al., 2009) – has become widespread in health care (Schouten et al., 2008) and is now moving into other social service sectors including education (Bryk et al., 2015; Hough et al., 2017). Traditionally, the primary method in continuous improvement for substantiating that changes are improvements is statistical process control (SPC). Under SPC, statistical tools, namely run charts (Perla et al., 2011) and control charts (Mohammed, 2004), are constructed in the field and then examined using a set of decision rules to provide evidence for a warrant that changes made led to improved quality and should be pursued further. A strong, timely warrant is essential to the success of continuous improvement (Reed & Card, 2016).

Run charts and control charts are enhanced point and line plots of a signal over time accompanied by a diagnostic test for nonrandom variation. The plot portion typically includes the signal data, a center line, and control limits for the signal. For control charts, the center line is the expected mean, and the upper and lower control limits are three standard deviations above and below the mean, respectively. The run chart uses the expected median for the center line, to reduce the influence of outlying observations, and does not include control limits. To substantiate improvement, a diagnostic test consisting of one or more decision rules is applied to the chart. For control charts, a common decision rule asks if any points in the signal fall outside of the control limits (Anhøj & Wentzel-Larsen, 2018). Run chart decision rules

generally ask if any runs – sequential points on one side of the median – are longer than would be expected (Anhøj & Wentzel-Larsen, 2018). If any of the decision rules for a chart test positive, then there is evidence that the signal is changing.

Although there has been some innovation in SPC recently (see De Ketelaere et al., 2016; Ferrer, 2014), the run charts and control charts used in industry and health care today largely resemble the tools proposed by Shewhart (1931) at Bell Laboratories almost a century ago. This is, in part, a tribute to the strength of the conceptual underpinnings of SPC. In SPC, processes are considered either "in control" or "out of control." For a process "in control," the future can be predicted by looking at the past. This is because all variation in a controlled process arises from what Shewhart (1931) referred to as common causes, i.e., expected random sources. Consequently, a controlled process is also an ergodic process with exchangeable data. In practice, this means the distributions of measures associated with the process can be modeled stochastically and represented with control limits – boundaries within which future data will fall with some probability (Shewhart, 1931).

Shewhart (1931) may not have had the same language at the time. But I would contend that SPC, by today's standards, is a rigorous causal method. Similar methods like single-subject case designs can meet What Works Clearinghouse standards without reservations and can be considered strong evidence under ESSA given sufficient spread and scale (What Works Clearinghouse, 2020). When the assumptions are met, SPC can provide evidence as strong as a Randomized Controlled Trial (RCT) because SPC is an experimental methodology. The difference between SPC and an RCT is the counterfactual which in SPC is estimated from historical data instead of data from a comparison group. Shewhart (1931) considered gaining control of a

process to be the goal of manufacturing because once control had been established then changes in the process could be detected.

Despite the conceptual strength of SPC as proposed by Shewhart (1931), the tools of SPC, run charts and control charts, have some important limitations. First, run charts and control charts only work with serially independent data. If the process has a temporal element which results in autocorrelation, seasonality, or a trend then data from the process must be modeled before SPC can be performed on the residuals (Box & Narasimhan, 2010; Superville & Adams, 1994). This limits the applications where run charts and control charts alone are effective. For example, SPC performed worse than Twitter's anomaly and breakout detection algorithm when both were applied to data on health care-associated infections exhibiting autocorrelation and seasonality (Wiemken et al., 2018).

Second, the decision rules for run charts and control charts recommend collecting at least 10 data points and preferably 20-30 data points (Anhøj & Wentzel-Larsen, 2018; Langley et al., 2009). This means run charts and control charts are difficult to use effectively with data captured weekly or more infrequently. In industry, where processes are often cycled in minutes or hours this is not a problem. However, when processes are cycled in weeks or even months this can be a serious limitation. Health care has some processes with short timelines. However, the targets for improvement in health care often report data on a much slower cadence. In a systematic review of health care's use of the Plan-Do-Study-Act cycle, an iterative improvement methodology, only 14% of studies used data collected monthly or more frequently (Taylor et al., 2014). Although the review suggests that the infrequent data collection is the result of

improper improvement methods, the results may also indicate that improvable targets in health care are often on a longer time scale.

Despite these limitations, run charts and control charts are still used widely in improvement work. This may be, in part, because methodological limitations are often downplayed in practice. But importantly, it is also because methodological advances have been slow to be accepted into SPC (Woodall, 2000). Woodall (2000) attributes this reticence to a number of reasons including the weighty bureaucracy of quality assurance, the familiarity and ease of existing methods among quality professionals, weak statistical backgrounds in the field, and the underdevelopment of an organized and tested literature base. In sum, methodological advances over the past century have not been useful enough to justify a change in the entrenched field of quality improvement, especially when compared to proven methods like run charts and control charts which are straightforward and easy to use.

In the present study, I introduce the Dynamic Control Chart (https://github.com/westdew/dccharts/) – a novel diagnostic test for substantiating improvement that is based on statistical techniques designed for making causal claims using automated short-term forecasts (Brodersen et al., 2015). Recently, with the proliferation of big data, methodologies for complex automated analyses of time-series data have grown in popularity. These methodologies address analytic decision making either empirically or heuristically to reduce the cost associated with conducting thousands of analyses directly. For instance, automated short-term forecasting, also called nowcasting, has been used by online advertising providers to present clients with timely impact estimates on advertising campaigns (Brodersen et al., 2015; Scott & Varian, 2014). Dynamic control charts will be presented as a

simple point-and-line plot with an accompanying diagnostic test available in an R package that mirrors the experience of using run charts and control charts. Behind this facade is an automated method based on the R package Causal Impact (Brodersen et al., 2015) which fuses state-space models (Durbin & Koopman, 2012) and Bayesian automatic variable selection (George & McCulloch, 1997) from applied statistics with synthetic control methods from Political Science's quantitative comparative case-studies (Abadie et al., 2015) to make causal claims using automated short-term forecasts (Scott & Varian, 2014).

Dynamic control charts are a novel approach for substantiating improvement for several reasons. First, to the knowledge of the author, dynamic control charts are the first fully automated diagnostic test for continuous improvement. There are many tools for constructing the statistical plots used in traditional SPC. However, the interpretation of those plots requires, at a minimum, the ability to read the plot and apply an appropriate set of decision rules, e.g., the Western Electric rules for Shewhart control charts (Anhøj & Wentzel-Larsen, 2018). A pure diagnostic test gives the user a simple positive or negative result. In the case of continuous improvement, a positive result for a dynamic control chart would provide evidence that a change was an improvement. Together with qualitative impressions from improvers and an interpretation of the point-and-line plot of the data this evidence could support a warrant to pursue the change further.

In addition to the novelty of a fully automated diagnostic test, dynamic control charts are also potentially a much stronger diagnostic test than the decision rules used in traditional SPC. Run chart and control chart decision rules are intentionally simplistic to support their ease of use, especially in the field. However, their necessary simplicity underutilizes the available

data. The decision rules for run charts, for instance, are based on the length of runs, i.e., consecutive points, on either side of the expected median (Z. Chen, 2010; Schilling, 2012). If there are longer runs than would be statistically likely then the data may have changed. This rule in no way considers how far the data are from the median. Large effects are detectable quicker by visual inspection than with a run chart because run charts don't give any extra weight to the size of the effect. Even the detection of small effects is greatly hindered because the runs rules don't accumulate the slight deviation of each point from its counterfactual. By ignoring this information, the strength of the run chart diagnostic test is weakened. Similarly, the decision rules for control charts rely on how the data fall in relation to standard deviations of variance from the mean (Anhøj & Wentzel-Larsen, 2018). The rules for control charts are more complicated and account for more of the information contained in the data. However, they still eschew available information for the sake of simplicity. The result likely is more data are required by run charts and control charts to have a strong test. Dynamic control charts are a Bayesian model-based approach for both the construction of the counterfactual and the subsequent diagnostic test. This means dynamic control charts are potentially a stronger diagnostic test which would require less data to reach similar levels of performance.

Another novel aspect of dynamic control charts is the approach taken in addressing temporal dynamicity, e.g., autocorrelation and nonstationarity. Statistical monitoring of processes that naturally vary over time (e.g., chicken egg production, see Mertens et al., 2009) is typically accomplished with synergistic control, whereby statistical process control is performed on the residuals of another statistical model, e.g., an ARIMA model (Box & Narasimhan, 2010; De Ketelaere et al., 2016; Superville & Adams, 1994). Dynamic control

charts, on the other hand, use a state-space model which can simultaneously estimate the dynamic temporal state and the underlying stable process distribution (Scott & Varian, 2014). This approach has at least three clear advantages over synergistic control. First, estimating a traditional statistical model, e.g., an ARIMA model, is not an easily automatable task, as it is highly parametric. The models used in dynamic control charts were built to be automated and include important features that make automation possible like Bayesian automatic variable selection (George & McCulloch, 1997). If dynamicity is present in the data, dynamic control charts can be easily adjusted to attempt to accommodate these features. Second, forms of nonstationarity (e.g., trends, drifts, cycles, seasons) are not handled well by traditional statistical models. For instance, in an ARIMA model non-stationary data is repeatedly differenced until it becomes stationary which has the undesirable effect of both losing information and inducing a problematic phenomenon called forecast recovery where shifts in the data induce impulse response functions, i.e., decaying pulses (Q. Chen et al., 2009). Third, state-space models can be expanded to include other regressors, e.g., controls, and, using Bayesian automatic variable selection, these regressors can be empirically culled to the optimal predictive set. In sum, the statistical methods behind dynamic control charts have the potential to accommodate many different forms of temporal dynamicity, often in an automated fashion. This advantage strongly differentiates dynamic control charts from existing methods like synergistic control where automation is not possible.

In the present study, I address the aforementioned claim that the Dynamic Control Chart is a much stronger diagnostic test than traditional alternatives. To do so, I evaluate and compare the diagnostic value of dynamic control charts, run charts with Anhøj decision rules

(Anhøj, 2015), and Shewhart Control Charts with Western Electric decision rules (Anhøj & Wentzel-Larsen, 2018). Dynamic control charts are hypothesized to be a superior test for detecting change in normally distributed exchangeable data and normally distributed autocorrelated data. In particular, dynamic controls charts are expected to greatly outperform current alternatives when there is limited data available. The performance of dynamic control charts will, first, be examined for the base automated dynamic control chart which only includes a static intercept in the counterfactual model. Then, the performance will also be examined for a more complicated model including an autoregressive component. The congruity of the model to the data is expected to influence the effectiveness of dynamic control charts. However, dynamic control charts are still hypothesized to be a superior test for detecting change even when the model-data pairing is incongruous.

Methods

To test the hypotheses of the present study, I will use a similar method to that employed by Anhøj in examining the diagnostic value of the decision rules for run charts (Anhøj, 2015) and control charts (Anhøj & Wentzel-Larsen, 2018). First, I will simulate appropriate time-series data both with and without a shift in the process mean, i.e., non-random variation. Then, I will apply dynamic control charts, run charts with Anhøj decision rules, and Shewhart control charts with Western Electric decision rules to these time-series data sets. For each simulation, I will know if there was a non-random shift in the process mean and whether each SPC method identified the simulation as potentially containing a non-random shift. Finally, I will use likelihood ratios – a common statistic for examining diagnostic

performance in health care – to evaluate how well the SPC methods diagnosed non-random variation in the data. Likelihood ratios will be compared across different SPC methods and different forms of data to provide evidence for the study's hypotheses.

Simulation of Data

There are many forms of data used in improvement work. Much of the complexity of control charts, for instance, arises from matching the data distribution to the correct control chart. In the present study, I will be limiting my analyses to continuous data which can be examined with dynamic control charts (using an identity link function), run charts, and Shewhart's I-Chart (used for monitoring individuals). Data from discrete distributions can also be examined with dynamic control charts by using a proper link function. However, this type of analysis is outside of the scope of this work.

The present study's hypotheses investigate change detection in two forms of continuous data: exchangeable data and autocorrelated data. I used the R programming language (R Core Team, 2020) to simulate all study data. Exchangeable data from an ergodic process, by definition, is drawn randomly from a fixed distribution. For the present study, I simulated exchangeable data using a normal distribution with a fixed mean (M = 0) and standard deviation (SD = 1). Autocorrelated data cannot be drawn randomly from a distribution as the sequencing is important. For the present study, I used an ARIMA model to simulate autocorrelated data with various levels of autoregression using innovations drawn from a normal distribution with a fixed mean (M = 0) and standard deviation (SD = 1). The precise levels of autoregression simulated – low (AR1 = 0.2) and medium-low (AR1 = 0.4) – were chosen

such that the autocorrelation would be substantial enough to impact the results but not so large as to be immediately visible upon inspection (Stoumbos & Reynolds, 2000). Additionally, supplemental analyses of highly (AR1 = 0.8) autocorrelated data demonstrated very poor performance across SPC methods.

For all analyses, I simulated thousands of appropriate data sets (N = 10000) which were representative of time-series data captured from an imagined 'in control' process. The number of simulated data sets was chosen based on prior studies of SPC in the literature (Anhøj, 2015; Anhøj & Wentzel-Larsen, 2018). In half of the data sets in each of the analyses (N = 5000), I introduced a non-random mean shift of two standard deviations (D = 2) after some fixed number of data points. This mean shift was representative of an abrupt change in the underlying process. The functional form and magnitude of the change were chosen based on prior studies of SPC in the literature (Anhøj, 2015; Anhøj & Wentzel-Larsen, 2018). The simulated data sets which included a non-random shift were used in analyses to detect rates of true positives and false negatives. Then, the other half of the simulated data sets without a shift were used in analyses to detect rates of true negatives and false positives.

Another important consideration in the simulation of the data was the number of data points included for calibration (hereafter referred to as the pre period) and detection (hereafter referred to as the post period). In a prior study of run charts in the literature combinations of 6, 12, and 18 pre and post data points were tested (Anhøj, 2015). However, in a different study including control charts 10, 20, and 40 data points were tested (Anhøj & Wentzel-Larsen, 2018). This is most likely because control charts require more data than run charts. I chose to simulate combinations of 6, 12, and 24 pre and post data points as a compromise which would contrast

the performance of control charts when there was limited data available. Recall, I had hypothesized that dynamic control charts would demonstrate the largest performance difference when data was sparse. In total, I simulated banks of data sets for normally distributed exchangeable data and multiple strengths of autocorrelated data – low (AR1 = 0.2) and medium-low (AR1 = 0.4) – for all combinations of 6, 12, and 24 pre and post data points.

Tested Methods

For each bank (27 in total) of simulated data sets (N = 10000), I applied the four SPC methods under review: run charts with Anhøj decision rules, Shewhart control charts with Western Electric decision rules, dynamic control charts with a static intercept, and dynamic control charts with an autoregressive component. The SPC methods each independently diagnosed whether non-random variation may have been present for each simulated data set. Since I also simulated the data, I knew whether non-random variation, in the form of a mean shift, was truly present. When the SPC method's diagnosis was correct a true positive or true negative was observed. Conversely, when the diagnosis was incorrect a false positive or false negative was observed. For each analysis of each bank of simulated data sets, a confusion matrix was assembled by counting up the records of true positives, false negatives, false positives, and true negatives. In total, I examined the diagnostic value of 4 SPC methods on 3 forms of data for 3 possible pre period lengths and 3 possible post period lengths. This produced 108 (4 x 27) confusion matrices. In the sections that follow, I explain the use of each SPC method in greater detail before addressing the statistic, likelihood ratios, used to compare the results.

Run Charts with Anhøj Decision Rules

The run chart is a point and line plot of a signal over time enhanced by the inclusion of the expected median. Run charts rely on the statistics of runs – sequential points on one side of the median – to detect non-random variation (Z. Chen, 2010; Schilling, 2012). For example, the Anhøj decision rules for run charts asks two questions about runs (Anhøj, 2015): 1) Is the number of runs less than would be statistically expected given the quantity of data? 2) Is the longest run beyond what would be statistically expected given the quantity of data? If either decision rule tests positive, then there is evidence that the process is changing. The only assumption made by the run chart is that the data is exchangeable.

In the present study, run charts were produced using the R package qicharts2 (Anhøj, 2018). Figure 3.1 presents an example of a Run Chart created by the package qicharts2. For each run chart, qicharts2 calculates the median during the pre period and then reports the number of times the signal crosses the median, i.e., the number of runs, and the length of the longest run in the post period. These statistics were used with the Anhøj decision rules (Anhøj & Wentzel-Larsen, 2018) to diagnose if there was non-random change in the post period. The Anhøj decision rules specify the smallest number of crossings expected and the longest run expected based on the number of data points in the post period, n:

$$\begin{aligned} \text{Upper limit for longest run} &= \text{round}(log_2(n) + 3) \\ \text{Lower limit for number of crossings} &= \text{qbinom}(0.05, n-1, 0.5) = k, \\ &\quad \text{where } Pr(B(n-1, 0.5) = k) = 0.05 \end{aligned}$$

If either signal is beyond the limit, then non-random variation may be present in the post period. For example, if the post period contains 10 points, random variation should result in no

less than 2 crossings and no more than 6 points in a run. If data is observed with 0 or 1 crossings or a run of 7 or more points, then non-random variation may be present.

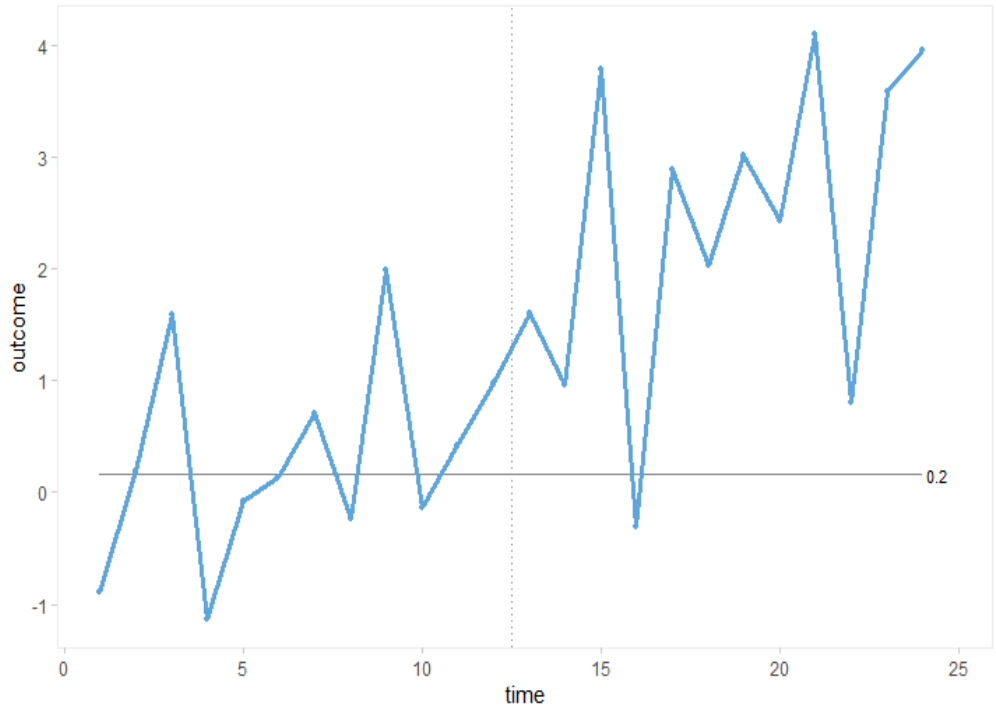

Figure 3.1. An example of a Run Chart where non-random change was present and detected by Anhøj decision rules. The blue line graph is the raw outcome data (M=0, SD=1). The dashed line is the beginning of the non-random mean shift (D=2). The grey line is the median of the pre period.

Shewhart Control Charts with Western Electric Decision Rules

A control chart is an enhanced point and line plot of a signal over time. The plot portion typically includes the signal data, a center line, and control limits for the signal. Traditionally, the center line is the expected mean, and the upper and lower control limits are three standard deviations above and below the mean, respectively. To substantiate improvement, decision rules – tests which compare the signal data to the control limits – are applied to the control chart. For example, the Western Electric decision rules ask four questions: 1) Do any points in

the signal fall outside of the control limits? 2) Do any two out of three points fall outside of two-thirds of the control limits? 3) Do any four out of five points fall outside of one-third of the control limits? 4) Do any eight successive points fall on one side of the center line? If any of the decision rules test positive, then there is evidence that the process is changing. The control chart makes two important assumptions: it assumes the data is exchangeable and the distribution of the data is well-defined. Though, control charts have been shown to be robust to violations of the second assumption (Stoumbos & Reynolds, 2000).

Shewhart I-Charts were produced using the R package qicharts2 (Anhøj, 2018). Figure 3.2 presents an example of a Shewhart I-Chart created by the package qicharts2. For each control chart, qicharts2 calculates the mean as well as the upper and lower control limits based on the data in the pre period. The upper and lower control limits are 3 standard deviations (3-sigma) above and below the mean, respectively. These statistics were then used with Western Electric (WE) decision rules 1-3 (Anhøj & Wentzel-Larsen, 2018) to diagnose if there was non-random change in the post period. The 4th WE decision rule was excluded because it requires a minimum of eight data points in the post period, and my hypotheses expressly aimed to identify the diagnostic value of SPC methods when there is minimal data available. Anhøj & Wentzel-Larsen (2018) found only minimal difference in the diagnostic value of the Western Electric decision rules when omitting the 4th rule.

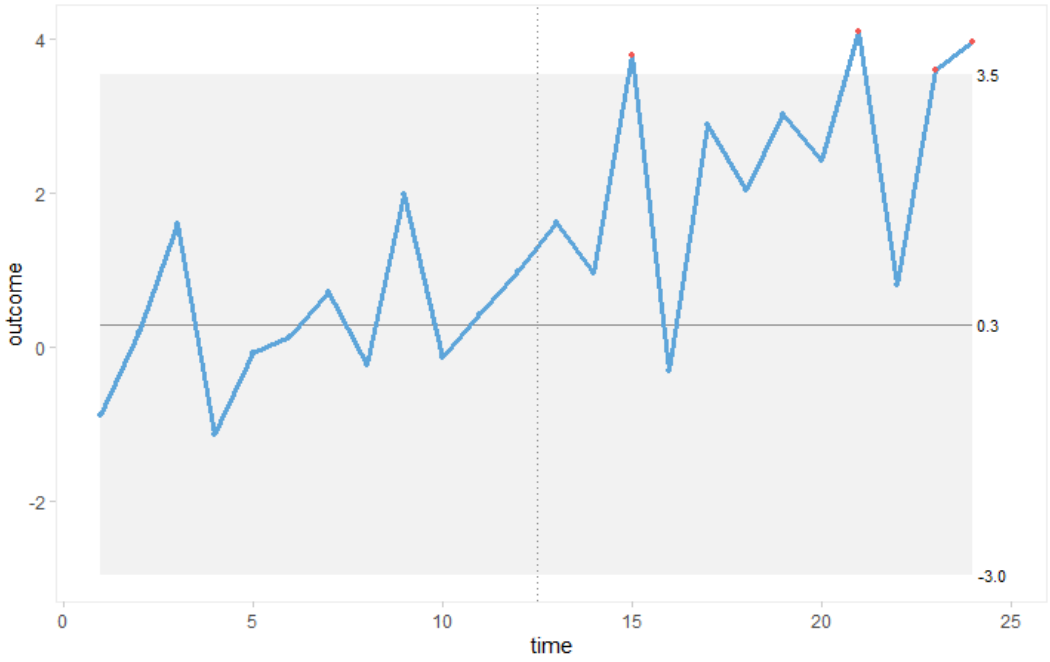

Figure 3.2. An example of a Shewhart I-chart where non-random change was present and detected by WE decision rules 1-3. The blue line graph is the raw outcome data (M=0, SD=1). The dashed line is the beginning of the non-random mean shift (D=2). The grey line is the mean of the pre period. The grey ribbon is the region bounded by the 3-sigma control limits.

The WE decision rules specify the minimum rate of data beyond the control limit which indicates non-random variation. If this rate is observed in the data in any period, then the test is positive. Likewise, if the data beyond the control limit never reaches this rate, then the test is negative. If any of the decision rules tests positive, then non-random variation may be present in the post period. The WE decision rules 1-3 are as follows:

1) One or more points beyond the 3-sigma control limits.

2) Two out of any three points beyond 2-sigma (two thirds of the control limits).

3) Four out of any five points beyond 1-sigma (one third of the control limits).

Note, control chart decision rules (presented above and in the literature) do not take into account the number of points in the post period. Consequently, as the post period lengthens there will be more false positives and less false negatives. This will increase the sensitivity and decrease the specificity of the diagnostic.

Dynamic Control Charts

Dynamic control charts are an automated method based on the R package Causal Impact (Brodersen et al., 2015) which fuses state-space models (Durbin & Koopman, 2012) and Bayesian automatic variable selection (George & McCulloch, 1997) from applied statistics with synthetic control methods from Political Science's quantitative comparative case-studies (Abadie et al., 2015) to make causal claims using automated short-term forecasts (Scott & Varian, 2014). Dynamic control charts were produced using the R package dccharts (https://github.com/westdew/dccharts/). This package – introduced in the present study – provides an R interface for using Bayesian structural time-series (Scott & Varian, 2014) to substantiate improvement in a manor akin to traditional SPC.

The dccharts package performs three important functions. First, it assembles and estimates a state-space model using the package BSTS (Scott & Varian, 2014). Much like the R packages for statistical process control, dccharts only requires the selection of a model type (e.g., intercept), the provision of the signal data, and an indication of when the post period begins. Other BSTS parameters like the number of Gibbs samples drawn and the length of the burn-in period were given the default values suggested in the Causal Impact package (Brodersen et al., 2015). Second, dccharts generates the counterfactual estimate from the

state-space model using the technique specified in the Causal Impact package (Brodersen et al., 2015). Third, dccharts produces post estimation statistics (e.g., p-values), diagnostic results, and plots. Again, these were based on the open-source code base of the Causal Impact package (Brodersen et al., 2015).

An analyst could accomplish the same functions as dccharts using the package BSTS or Causal Impact. However, the use of BSTS would require many decisions on the analyst's part as it is a generic package for Bayesian estimation of state-space models, and the use of Causal Impact would limit the analyst's modeling approach unless a custom BSTS model is specified.[2] Consequently, dccharts was created to provide a package which had similar inputs and outputs to existing R packages for statistical process control while performing estimation for different model types using BSTS (Scott & Varian, 2014) and post estimation functions based on the techniques used in Causal Impact (Brodersen et al., 2015).

For each dynamic control chart in the present study, the dccharts package estimates a counterfactual for the post period based on the model chosen and the pre period data. Then, a p-value is calculated by examining where the actual post period data fell within the probability distribution of the estimated counterfactual. This Bayesian p-value indicates the proportion of the Bayesian simulations of the counterfactual where the data was more extreme than the actual post period data. It can be interpreted, similar to a frequentist p-value, as the probability of obtaining a result at least as extreme as the observed data, assuming the null hypothesis is true.

[2] By default, Causal Impact assumes a synthetic control method with multiple contemporaneous covariates and a local level state specification.

Dynamic control charts, then, compare this p-value to an alpha value or specificity threshold (referred to as the significance level in educational research) to determine if non-random variation may be present. This threshold can be tuned based on the diagnostic needs of the task. For example, if the analyst were more willing to accept more false positives because the evidence was from an exploratory study then the threshold could be increased. In the present study, for dynamic control charts with a static intercept the threshold was set to $\alpha = 0.01$, and for dynamic control charts with an autoregressive component the threshold was set to $\alpha = 0.025$.

As noted above, the present study's hypotheses test two types of dynamic control charts. The first type of control chart is the base automated dynamic control chart which includes only a static intercept. Figure 3.3 presents an example of a dynamic control chart with a static intercept created by the package dccharts. This type of control chart assumes that the data is normally distributed and exchangeable. The counterfactual is estimated using a state-space model defined by the following equations:

$$y_t = \alpha_t + \epsilon_t$$
$$\alpha_t = M$$
$$\epsilon_t \sim N(0, \sigma_{\epsilon_t}^2)$$

The first equation is referred to as the observation equation. In this equation, y_t is the outcome for the t^{th} observation, α_t is the underlying state for the t^{th} observation, and ϵ_t is the independent normally distributed errors in the measurement of the state. The underlying state is defined in the second equation where, for a static intercept model, M is a constant. The parameters estimated by this model are the mean of the pre period, M, and the variance of the

pre period, $\sigma^2_{\epsilon_t}$. All priors were specified as unassuming. A state-space model is not necessary to model a straightforward distribution like this. However, all models in the Bayesian Structural Time-series package are state-space models where the user customizes the model by assembling the state specification. This affords flexibility to Bayesian Structural Time-series models which can, consequently, easily include many different components.

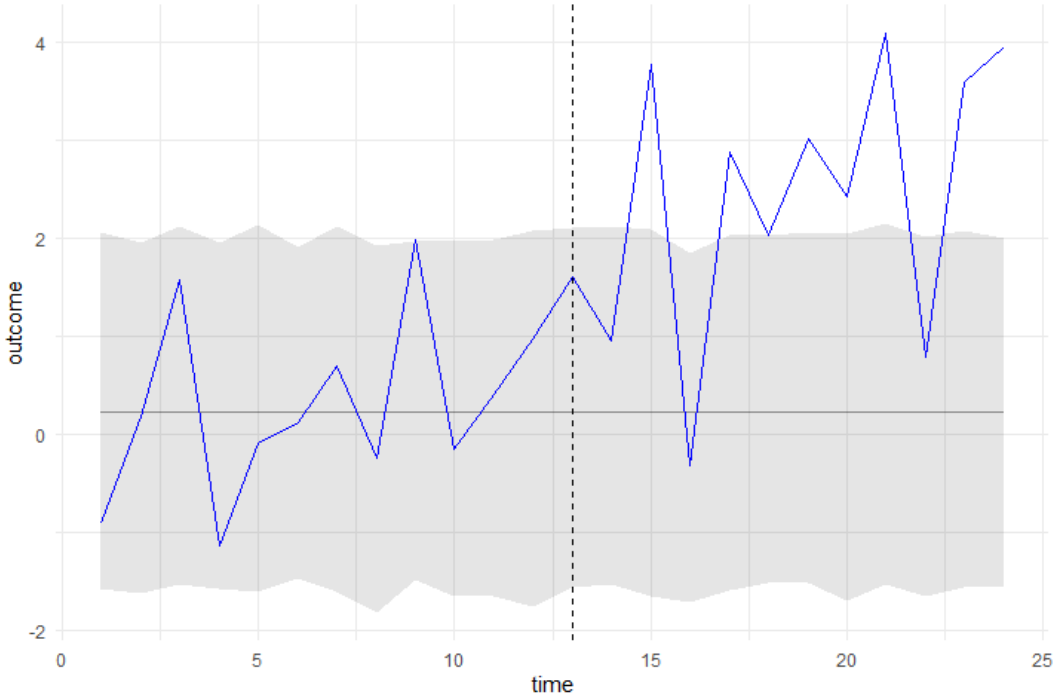

Figure 3.3. An example of a Dynamic Control Chart Intercept Model where non-random change was present and detected (p=0.0011). The blue line graph is the raw outcome data (M=0, SD=1). The dashed line is the beginning of the non-random mean shift (D=2). The grey line is the model fit to the left of the dashed line, and the counterfactual to the right of the dashed line. The grey ribbon is the 95% Bayesian credible interval.

The second type of control chart is a dynamic control chart with an autoregressive component. Figure 3.4 presents an example of a dynamic control chart with an autoregressive component created by the package dccharts. In this type of control chart, the counterfactual is estimated using the following state-space model:

$$y_t = \alpha_t + \epsilon_t$$
$$\alpha_t = M + \phi_1 \alpha_{t-1} + \cdots + \phi_9 \alpha_{t-9} + \epsilon_{t-1}$$
$$\epsilon_t \sim N(0, \sigma^2_{\epsilon_t})$$

In the observation equation, y_t is the outcome for the t^{th} observation, α_t is the underlying state for the t^{th} observation, and ϵ_t is the independent normally distributed errors in the measurement of the state. In the state equation, M is the constant intercept, ϕ_p is the AR coefficient of the p^{th} lag, α_{t-p} is the p^{th} prior observation's state where a total of nine lagged states are included, and ϵ_{t-1} is the independent normally distributed error in the prior state. The parameters estimated by this model are the static intercept, M, the nine AR coefficients, ϕ_p, and the variance of the state error, $\sigma^2_{\epsilon_t}$.

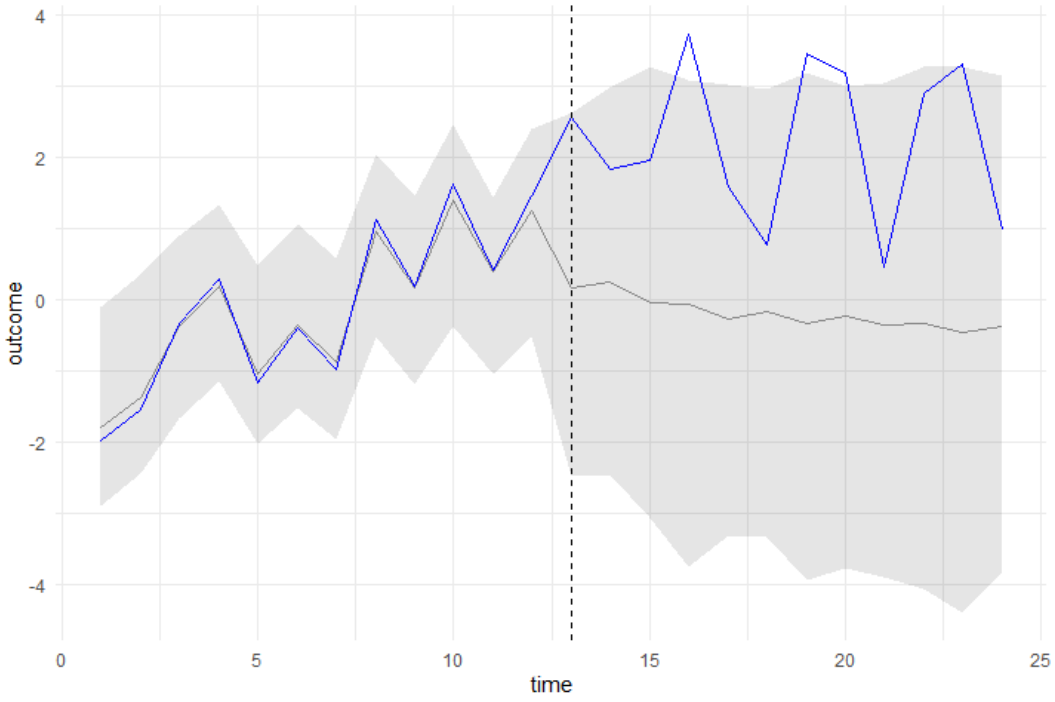

Figure 3.4. An example of a Dynamic Control Chart Autoregressive Model where non-random change was present and detected (p=0.01). The blue line graph is the raw outcome data (M=0, SD=1, Ar1=0.4). The dashed line is the beginning of the non-random mean shift (D=2). The grey line is the model fit to the left of the dashed line, and the counterfactual to the right of the dashed line. The grey ribbon is the 95% Bayesian credible interval.

One of the traditional problems when automating an AR model (a subset of the ARIMA family) is deciding the order of the model. The order determines the number of lagged coefficients to include. The model above is set up as an AR(9) model which would only be appropriate if the data justified nine significant AR parameters. However, the dynamic control chart uses a state specification referred to as an auto AR. This component has the same model specification as a simple AR model. However, when estimating the AR parameters, the estimates are encouraged to be zero if the coefficient is insignificant. This is accomplished by using a spike and slab prior for the Bayesian prior of the AR parameters (George & McCulloch, 1997). A spike and slab prior suggests that the value of a coefficient is most likely zero (i.e., a

spike at zero) but could be anything (i.e., a slab across all potential values). The spike, which is typically a Bernoulli prior, selects only the most promising subset of coefficients. In the case of an auto AR component, only the significant AR parameters are included. In fact, if there is no autoregression then the model will revert to just a static intercept.

Likelihood Ratios

To compare the diagnostic value of different SPC methods an appropriate metric is needed. When examining clinical diagnostic tests, the likelihood ratio is often the favored metric because it is practically relevant (Attia, 2003). The greater the value of the positive likelihood ratio, LR^+, the greater the probability the condition is present when a test is positive. Likewise, the smaller the value of the negative likelihood ratio, LR^-, the lower the probability the condition is present when a test is negative. Anhøj & Wentzel-Larsen (2018, p. 4) argues, "[traditional metrics like] sensitivity, and specificity are not that useful on their own – they describe how non-random variation predicts a signal, not how a signal predicts non-random variation, which is what we really want to know." For these reasons, the present study compares likelihood ratios when examining diagnostic value.

Positive and negative likelihood ratios were calculated using the following equations:

$$LR^+ = \frac{\text{true positive rate}}{\text{false positive rate}} = \frac{\text{sensitivity}}{1 - \text{specificity}}$$
$$LR^- = \frac{\text{false negative rate}}{\text{true negative rate}} = \frac{1 - \text{sensitivity}}{\text{specificity}}$$

The positive likelihood ratio can be interpreted as the increase of the likelihood that a change was an improvement when the SPC diagnostic test is positive; the negative likelihood ratio can

be interpreted as the decrease of the likelihood that a change was an improvement when the SPC diagnostic test is negative. Ideally, a diagnostic test should have a large positive likelihood ratio and a small negative likelihood ratio. In practice, there is typically a tradeoff between the two. In research, a statistical test with conventional power, $1 - \beta = 0.80$, and conventional significance, $\alpha = 0.05$, has a positive likelihood ratio of 16 and a negative likelihood ratio of 0.21. This is sensical for research, as false positives (i.e., incorrectly rejecting the null hypothesis) are often considered worse than false negatives in scientific fields. Put another way, scientists would rather miss a positive result than confirm an incorrect hypothesis. In health care, however, diagnostic tests often aim for the opposite in performance, preferring to incorrectly diagnosis the presence of a condition (false positives) rather than miss a curable illness (false negative).

In the present study, SPC methods will be compared to the standard benchmarks for diagnostic test performance from clinical practice (Anhøj & Wentzel-Larsen, 2018). A positive likelihood ratio greater than 10 will be considered strong evidence that the condition may be present, and a negative likelihood ratio less than 0.1 will be considered strong evidence that the condition may not be present. The condition being diagnosed in this study is the presence of a change in a signal.

Results

The positive and negative likelihood ratios from the simulations in the present study are displayed in Figure 3.5 and Figure 3.6. In Figure 3.5, the likelihood ratios for SPC methods tested on normally distributed exchangeable data are presented in a small multiple showing the

results for various lengths of pre data (rows) and post data (columns). Figure 3.6, then, presents a small multiple of likelihood ratios for SPC methods tested on autocorrelated data. The plot shows the results for various levels of autocorrelation (rows) and various lengths of post data (columns). Note, all results in Figure 3.6 had 12 points of pre data.

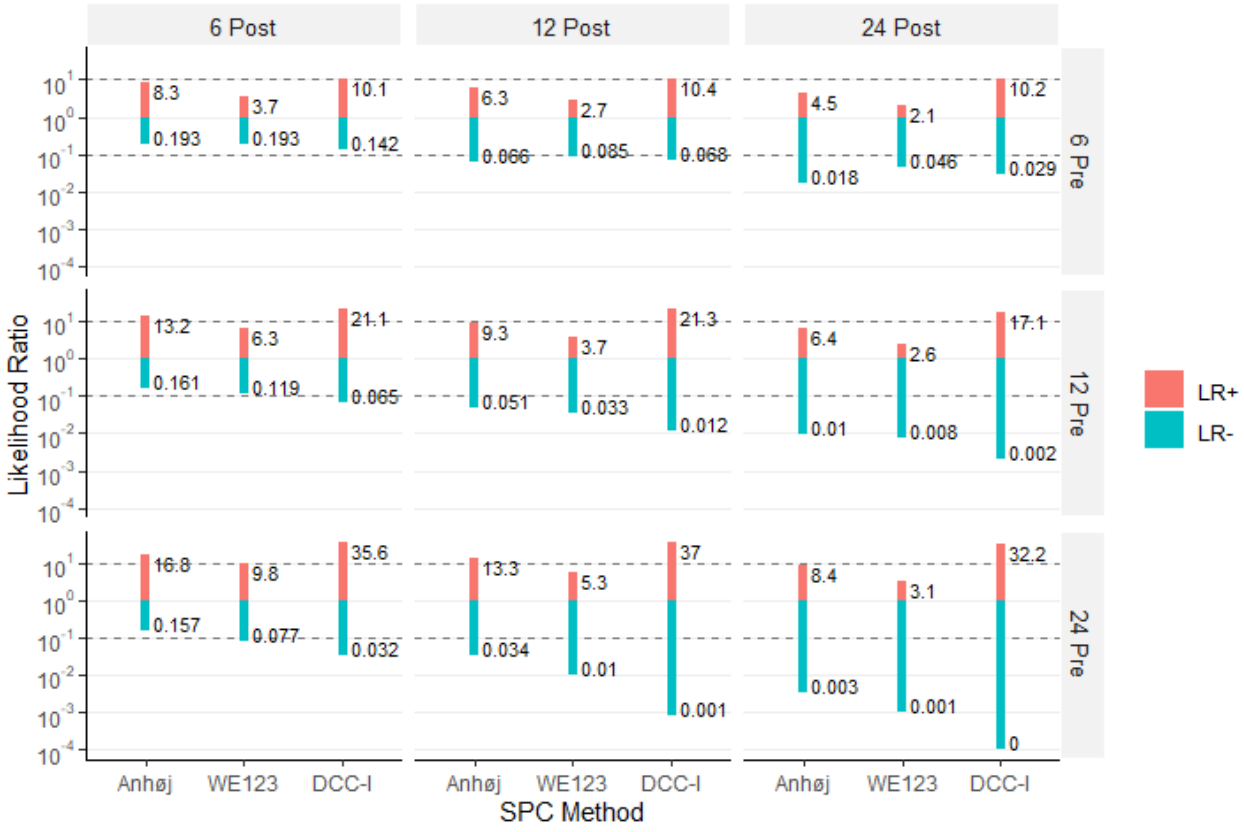

Figure 3.5. Likelihood ratios for SPC methods given normally distributed exchangeable data (M=0, SD=1) with various pre and post period lengths.

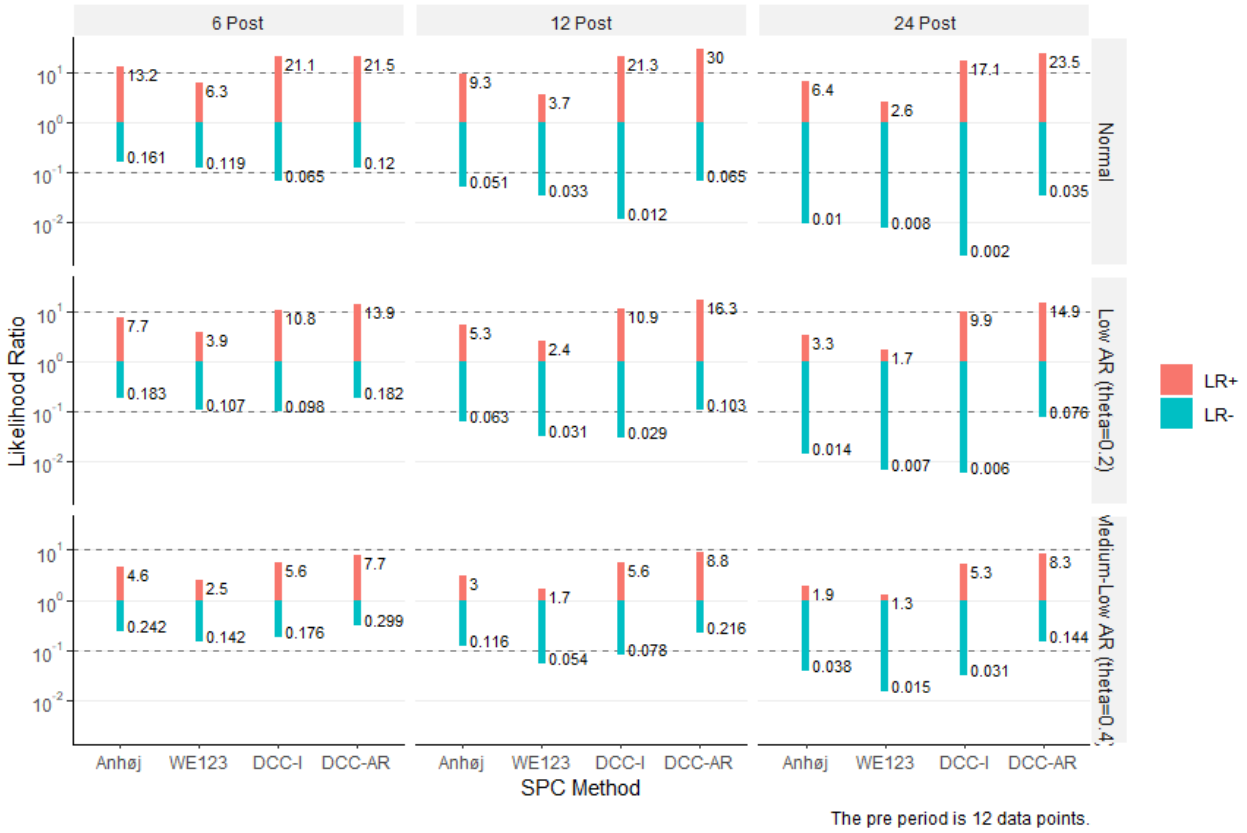

Figure 3.6. Likelihood ratios for SPC methods given autocorrelated data with normally distributed innovations (M=0, SD=1) and with various post period lengths and levels of autoregression.

Tests on Exchangeable Data

Overall, the dynamic control charts (DCC-I & DCC-AR) both performed better than traditional SPC methods in simulations of exchangeable data. Across all pre and post lengths, the dynamic control chart intercept model (DCC-I) displayed equivalent or larger likelihood ratios when compared to the run chart with Anhøj decision rules (Anhøj) and the Shewhart I-Chart with Western Electric decision rules (WE123). Moreover, the dynamic control chart intercept model (DCC-I) met the benchmarks for strong evidence ($LR^+ > 10$ and $LR^- < 0.1$) for all lengths of pre and post data except for the combination of 6 pre and 6 post points. In

contrast, traditional SPC methods only met this standard in one instance. The dynamic control chart autoregressive model (DCC-AR) also met the benchmarks for strong evidence and matched or exceeded the performance of traditional SPC methods. However, for exchangeable data, it did not outperform the dynamic control chart intercept model (DCC-I) which likely benefited from the congruous model-data pairing. In summary, for exchangeable data, the DCC-I was the best performer followed closely by the DCC-AR.

Figure 3.5 supports two potential reasons why traditional SPC methods only rarely met the benchmarks for strong evidence ($LR^+ > 10$ and $LR^- < 0.1$). First, the length of the pre period data appears to directly limit the strength of the diagnostic for all observed diagnostic tests. Increasing the length of the pre period increased the positive and negative likelihood ratios for every SPC method with every post period length. Consequentially, for shorter length pre periods the traditional SPC methods may not have had the strength to meet the benchmarks. The second reason why traditional SPC methods may have not met the benchmarks was the influence of the post period length on the balance of sensitivity and specificity. For all tested SPC methods, increasing the length of the post period increased the negative likelihood ratio. This most likely occurred because the additional data reduced the false negative rate. However, traditional SPC methods also saw a concomitant decrease in the positive likelihood ratio, i.e., an increase in the false positive rate. This tradeoff was particularly true for the Shewhart I-chart which does not consider the post period length in its decision rules. The dynamic control chart, however, was able to maintain its specificity as additional post period data was added. Ultimately, only a run chart with Anhøj decision rules (Anhøj), sufficient

pre period data (24 points), and a balanced quantity of post period data (12 points) was able to meet the benchmarks under the conditions of the study.

Tests on Autocorrelated Data

The dynamic control chart intercept model (DCC-I) also performed better than traditional SPC methods for detecting change in autocorrelated data. At low levels of autoregression (AR1 = 0.2), the intercept model had consistently higher likelihood ratios compared to the run chart with Anhøj decision rules (Anhøj) and the Shewhart I-Chart with Western Electric decision rules (WE123). Additionally, all these likelihood ratios met the benchmark for strong evidence. However, at medium-low levels of autoregression (AR1 = 0.4) the results were less conclusive. The dynamic control chart intercept model (DCC-I) had stronger positive likelihood ratios while traditional SPC methods had stronger negative likelihood ratios. In truth, the performance of all SPC methods without an autoregressive component were well below standards of strong evidence when diagnosing changes in data with medium-low levels of autoregression.

Unexpectedly, the dynamic control chart autoregressive model (DCC-AR) did not meet hypothesized expectations for autocorrelated data. For low levels of autoregression (AR1 = 0.2), the autoregressive model had consistently higher likelihood ratios compared to the run chart with Anhøj decision rules (Anhøj) and the Shewhart I-Chart with Western Electric decision rules (WE123). However, the dynamic control chart autoregressive model (DCC-AR) was not able to match the performance of the dynamic control chart intercept model (DCC-I). Furthermore, at medium-low levels of autogregression (AR1 = 0.4), the dynamic control chart autoregressive

model (DCC-AR) did not demonstrate performance that was consistently better than other SPC methods. When there was ample post period data (24 points), the dynamic control chart autoregressive model (DCC-AR) had the strongest and most balanced performance of the methods applied to autocorrelated data. But it did not meet the benchmarks, and when the post period only contained 6 or 12 points, the performance was well below standards of strong evidence. In summary, for weakly autcorrelated data, the DCC-I was still the best performer, and for higher levels of autocorrelation, none of the methods stood out or even performed adequately.

Discussion

The present study tested whether the Dynamic Control Chart was a stronger diagnostic test than traditional alternatives. My results confidently support this hypothesis. For normally distributed exchangeable data and weakly autocorrelated data, the base dynamic control chart with a static intercept performed substantially better than run charts and control charts under all tested circumstances. The Dynamic Control Chart demonstrated two important features which enabled its strong performance. First, the Dynamic Control Chart maintained a high positive likelihood ratio even as additional data was added to the post period. This strength can be attributed to the model-based approach which adapted well to variable post period lengths and maintained high specificity through a fixed specificity threshold, $\alpha = 0.01$. Second, the Dynamic Control Chart utilized the data to generate predictions which were more accurate than other SPC methods. For the simulations of exchangeable data, dynamic control charts accurately predicted the counterfactual for 95% of the data sets tested. Whereas, run charts

and control charts were 91% and 84% accurate, respectively. In sum, the Dynamic Control Chart was the best diagnostic test for detecting change in the simulations studied.

The present study also examined the performance of the autoregressive component for dynamic control charts. My hypothesis was that the effectiveness would be influenced by the congruity of the model-data pairing. Results, though, were inconclusive. For the lengths of pre and post data tested, dynamic control charts with an autoregressive component did not definitively outperform dynamic control charts with a simple static intercept even in autocorrelated data. Future research should examine why the autoregressive model was inferior to the static intercept model. The present studies results and methods suggest three possible reasons. First, the autoregressive model may require more data than the present study supplied. There was some evidence of accelerating performance as the pre and post periods were lengthened. Second, it's possible the automated form of the autoregressive model failed to correctly match the autocorrelation, decreasing the strength of the test. Future research should examine this possibility. Third, it could be that dynamic components don't add much to SPC methods, as the dynamicity is often indistinguishable from the change the method seeks to detect. In which case, the best approach may be to model the underlying causes of the dynamicity. There is great potential for dynamic control charts in this space. The automatic autoregressive component was easily included for the present study's tests. But other more specified forms (e.g., trends, cycles, seasons) could also be implemented and automated when needed.

There were some clear limitations with the present study. Simulations are useful for comparing methods under controlled conditions. However, often in the field unexpected and

complicated conditions are encountered. The present study cannot say for certain that the diagnostic values and relationships observed in these simulations would replicate using data from real world scenarios. For example, in the present study, even data with modest autocorrelation shrunk performance differences across SPC methods and skewed sensitivity making comparisons difficult. The strength of the dynamic control chart in the present study is evident in the results. But the strength of this method in the field will be best confirmed through its effective use in practice. To that point, the evidence presented suggests that continuous improvement professionals should consider using the Dynamic Control Chart, especially under certain circumstances. In the following paragraphs, I identify two situations where dynamic control charts seem preferable.

First, when there is limited data available in the post period, the dynamic control chart stood out for maintaining the potential to detect strong evidence of change. Furthermore, the dynamic control chart can theoretically provide some evidence even when there are less than six points in the post period. The present study did not investigate this scenario as the minimum data required for the decisions rules of a run chart is six points. But there is no technical reason why a dynamic control chart cannot make a prediction with only one post period point. Evidence from the present study suggests dynamic control charts are preferable when circumstances require responsiveness or data collection is limited.

Second, the dynamic control chart also stood out for maintaining the strength of its positive likelihood ratio. As was noted above, in simulations, dynamic control charts had high positive likelihood ratios even under the longest post periods. Dynamic control charts were also robust in the simulations with autocorrelated data where they maintained higher positive

likelihood ratios despite the tendency of autocorrelation to trigger more false positives. Remember, the positive likelihood ratio is the increase of the likelihood that a change was an improvement when the SPC diagnostic test is positive. If this statistic gets too small, then a diagnostic test cannot be confirmatory and other forms of evidence are needed to support the warrant that a change was an improvement. The present study suggests dynamic control charts are preferable when circumstances require that SPC methods provide the majority of confirmatory evidence for a warrant, especially if the chance of false positives is heightened.

There are also circumstances where a run chart or control chart may be preferable. For example, the run chart could be at an advantage when the functional form of the data is unspecified, as it is only parameterized by the median and makes few assumptions regarding the distribution of the data. Model based approaches, on the other hand, make assumption about the functional form of data. The priors used in the Bayesian state space models in a dynamic control chart assume that the residuals are normally distributed. This may give run charts an advantage under certain circumstances. However, the present study's investigation of the performance of dynamic control charts in autocorrelated data (which tends to be platykurtic) as well as a supplemental investigation examining the performance of dynamic control charts under various levels of skewness suggests that the strengths of the model-based approach may outweigh model misspecification.

Conclusion

The present study demonstrated that the Dynamic Control Chart has the potential to serve the same role as existing SPC methods in substantiating change in continuous

improvement work. Moreover, the simulations presented suggest that the Dynamic Control Chart could be a superior diagnostic test in many circumstances. Professionals working in continuous improvement should consider introducing the Dynamic Control Chart into their practice. Run charts and control charts are widely employed in improvement because they are straightforward and easy to use. Dynamic Control Charts have the promise to fill the same role while also providing stronger confirmatory tests, supporting any length of pre or post data, and modelling dynamic data.

Chapter 4 – Receiver Operating Characteristic Analysis of Dynamic Control Charts: A Case Study of Monthly School Attendance

Introduction

Continuous improvement – a strategy for advancing organizational quality (Deming, 2000; Langley et al., 2009) – has become widespread in health care (Schouten et al., 2008) and is now moving into education and other social service sectors (Bryk et al., 2015; Hough et al., 2017). One of the challenges of this transition has been applying existing continuous improvement methods from industry to fundamentally different sectors like education (Yurkofsky et al., 2020; Chapter 2). In the present study, I focus on the crucial matter of establishing evidence of improvement (Reed & Card, 2016) which has little extant research in education (see Chapter 2 and Chapter 3). The primary methods in industry for accomplishing this task – run charts (Perla et al., 2011) and control charts (Mohammed, 2004) – have many problems when used in educational improvement work, as human social dynamics violate their core assumptions (e.g., exchangeable data), they require too much data for the slower cadence of educational processes, and they demand a statistical skill set less prevalent among educators (see Chapter 2).

The Dynamic Control Chart (https://github.com/westdew/dccharts/) has been proposed as an alternative method for substantiating improvement in educational contexts which could potentially address these challenges (see Chapter 3). Dynamic control charts are based on the R package Causal Impact (Brodersen et al., 2015) which fuses state-space models (Durbin &

Koopman, 2012) and Bayesian automatic variable selection (George & McCulloch, 1997) from applied statistics with synthetic control methods from Political Science's quantitative comparative case-studies (Abadie et al., 2015) to make causal claims using short-term forecasts (Scott & Varian, 2014). The dynamic control chart was designed to be a largely automated method for establishing evidence of improvement while reducing the need for users to have statistical skill (Brodersen et al., 2015). Dynamic control charts are also a much stronger diagnostic test than run charts and control charts and, consequently, require less data to detect improvement (see Chapter 3). And, importantly, dynamic control charts have the potential to accommodate many different forms of temporal dynamicity, a feature common in educational data (Koopmans, 2020).

In the present study, I address the latter of these three claims. Using a case study approach, I examine the ability of dynamic control charts to meet the needs of an imagined continuous improvement effort focused on increasing student attendance. This is a timely and important aim as chronic absenteeism is a critical issue in today's schools (Balfanz & Byrnes, 2012; García & Weiss, 2018). Attendance was chosen because it is a dynamical process characterized by complexity which manifests in the data as autocorrelation, nonstationarity, and fractality (Koopmans, 2016). Run charts and control charts of attendance data at many different time scales (e.g., daily, weekly, monthly) are likely biased due to these dynamics (see Chapter 2 and 3). The dynamic control chart has tools available to address dynamicity including stochastic components which nonparametrically model the dynamics and the linear regression of predictors which can potentially control for much of the dynamicity.

Although dynamic control charts have the tools to appropriately detect improvement in dynamic data, there is a latent question of whether dynamic control charts will be a strong enough test to detect realistic effects under dynamic conditions. Dynamic control charts performed poorly detecting a large mean shift in simulations of moderate to heavily autocorrelated data (see Chapter 3). The issue is, changes in a signal, i.e., improvements, are dynamics themselves. In ergodic data, the effect must simply be large in comparison to the variance for it to be statistically significant. Dynamic signals have an additional problem where the effect must also become large quick enough in comparison to the dynamics of the data. Otherwise, the method may mistake the effect for normal dynamicity. Basically, for data from a dynamical process the further into the future you go the larger an effect needs to be to achieve significance. This can, however, be mitigated by including contemporaneous or historical predictors as a form of synthetic control (Abadie et al., 2015).

The present study addresses the above question by performing receiver operating characteristic (ROC) curve analyses (Swets, 1988) for four different types of dynamic control charts detecting simulated effects in authentic attendance data. The dynamic control chart types were selected to demonstrate a broad sample of the tools available for attending to dynamicity. The analyzed charts include: a basic dynamic control chart with only a static intercept (i.e., the naive model), a dynamic control chart with a local level component (i.e., a random walk) which nonparametrically models first-order polynomial dynamics (Durbin & Koopman, 2012), a dynamic control chart including the school's previous year's attendance as a predictor, and a dynamic control chart including the mean contemporaneous attendance of similar schools as a predictor. ROC curve analyses are a method from signal detection theory

used to assess the accuracy of diagnostic systems (Swets, 1988). They have a history of usage in this capacity in engineering, medicine (Hanley & McNeil, 1982; Zweig & Campbell, 1993), psychology (Swets et al., 2000), and, more recently, education (Bowers & Zhou, 2019; Christ et al., 2013; D'Agostino et al., 2018). Statistics from the ROC analyses will be used to compare the performance of the different types of dynamic control charts both to benchmarks from the literature and to each other. The aim of the study is to show the potential of the dynamic control chart for substantiating improvement in data from dynamical processes.

Methods

For the present study, authentic school-level attendance data (from New York City) which already features dynamics including autocorrelation, nonstationarity, and fractality (Koopmans, 2016) was used as the base data. Since these schools do not have known exogenous improvements in their attendance rates, I simulated improvement in the form of a non-random shift in the data. Then, I applied the different types of dynamic control charts to thousands of randomly selected school-level attendance data sets both with and without the simulated shift. For each test, I recorded whether a non-random shift was introduced and the p-value of the dynamic control chart diagnostic test. This pairing has a standard relationship where a lower p-value indicates a higher likelihood of the presence of a non-random shift. Finally, I used receiver operating characteristic (ROC) curve analysis on the results of the simulation study to examine the accuracy of each type of dynamic control chart.

Daily Attendance Data

Each day at 4pm the New York City (NYC) Department of Education publishes the provisional attendance data for all traditional public schools in the city. When these reports are finalized, they are compiled into a historical record of school-level daily attendance consisting of each day's enrollment, absent student count, present student count, and released student count. This data has been recorded since the 2005/06 school year and is available to the public on NYC Open Data (NYC Department of Education, 2018). From this data the daily attendance percentage was calculated as the number of students present divided by the number of students enrolled on a given day multiplied by 100.

The present study used school daily attendance data for 1692 schools from 2012/13 to 2016/17. However, a small number of these schools had nonstandard calendars or very small enrollments which would make them uncharacteristic of a typical urban school. I excluded schools with less than 10 enrolled students, less than 100 school days, or greater than 200 school days. Then, I also required that a school be present in all five years of the data and only included the most recent four years so that the prior year of attendance could easily be used in the analyses. There were 1527 schools remaining which with four years of data provided 6108 school-level attendance data sets for testing.

Daily school-level attendance data was then aggregated to the calendar month to create monthly average attendance percentages. I chose to test the dynamic control charts at the monthly scale for three reasons. First, I would contend this is the scale where the strongest causal evidence will be needed in improvement work. District leaders and improvement workers will be looking for evidence that their changes across the school year moved key

outcomes (e.g., attendance) to justify further improvement work and the direction of that work. Evidence at the weekly or daily scale can show the efficacy of specific parts of a change idea over a brief time span but can't convey long term improvements as well as monthly data. The second reason the monthly scale was chosen was to highlight some of the more challenging dynamics in attendance data. At the daily level attendance is characterized by predictable outliers, e.g., half days (see Figure 4.1) and a consistent weekly cycle which is highest mid-week (see Figure 4.2). At the weekly level attendance is characterized primarily by a small, localized trend (see Figure 4.3). At the monthly level, though, attendance is characterized by nonstationarity in the form of complex seasonal effects (see Figure 4.4). The third reason the monthly scale was chosen was because this scale has the least available data as monthly data in schools typically is limited to ten or less data points given the length of the school year. One of the strengths of the dynamic control chart is working with limited data (see Chapter 3).

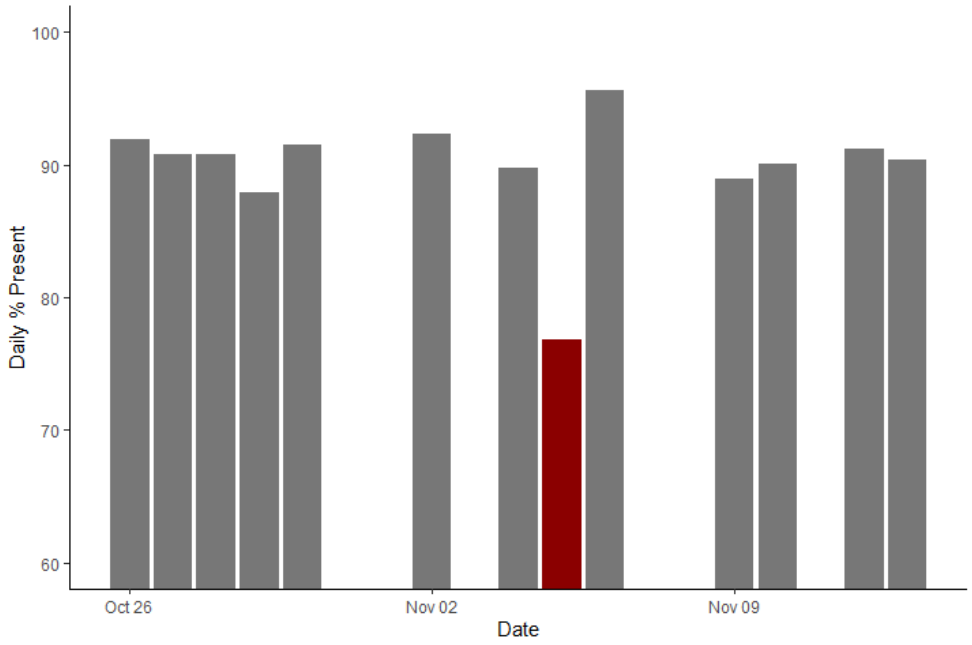

Figure 4.1. Example of an outlier in daily attendance data (red) which was discovered to be a half-day for parent teacher conferences.

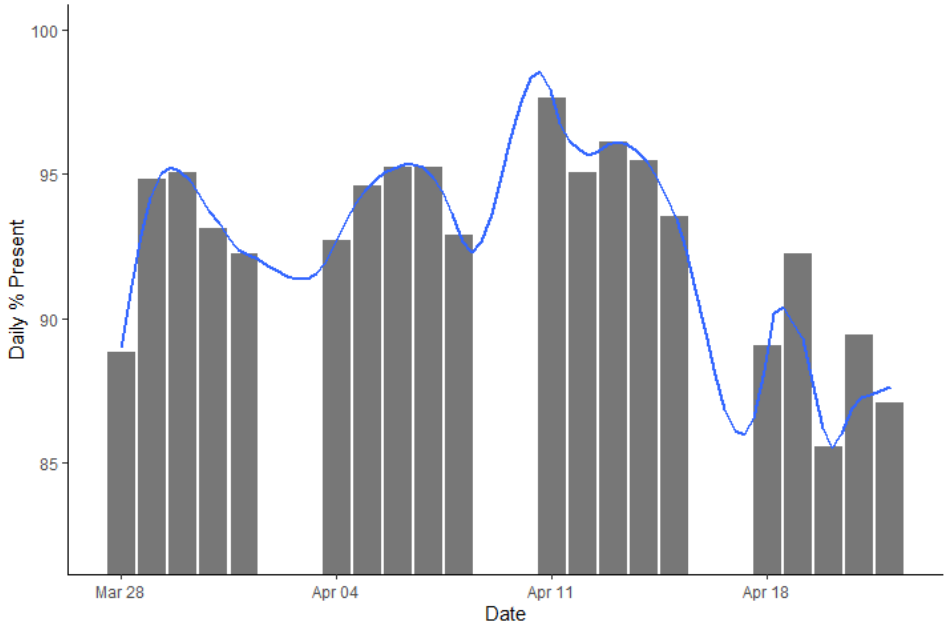

Figure 4.2. Example of a weekly cycle in daily attendance data where attendance in the middle of the week is higher than attendance at the ends of the week. Each group of five columns is a single week. The blue line is a loess curve selected to highlight the cycle.

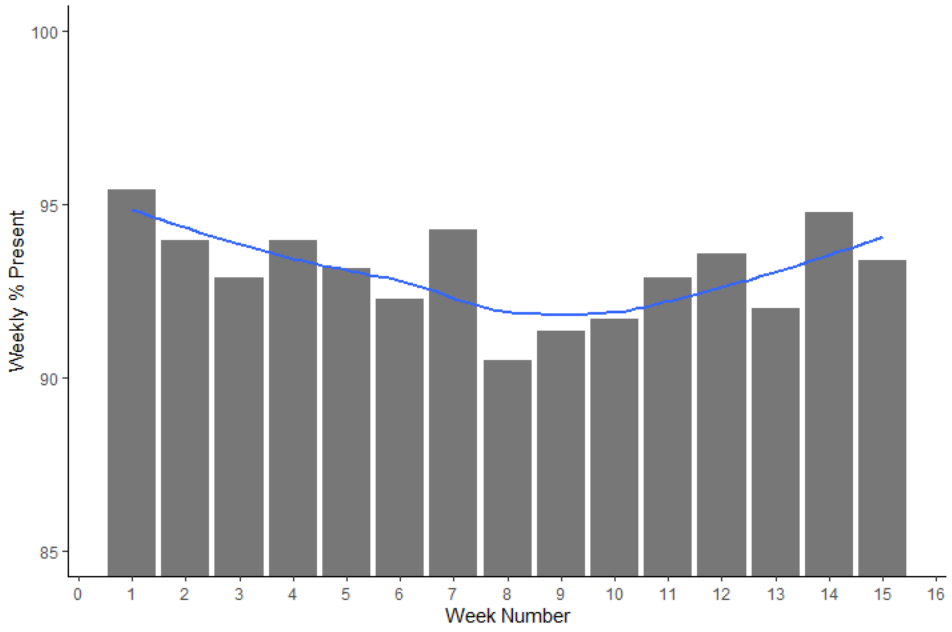

Figure 4.3. Example of localized trending present in weekly attendance data. From week 1 to 9 attendance is on a downward trend and from week 10 to 15 attendance is on an upward trend. The blue line is a loess curve selected to highlight the trending.

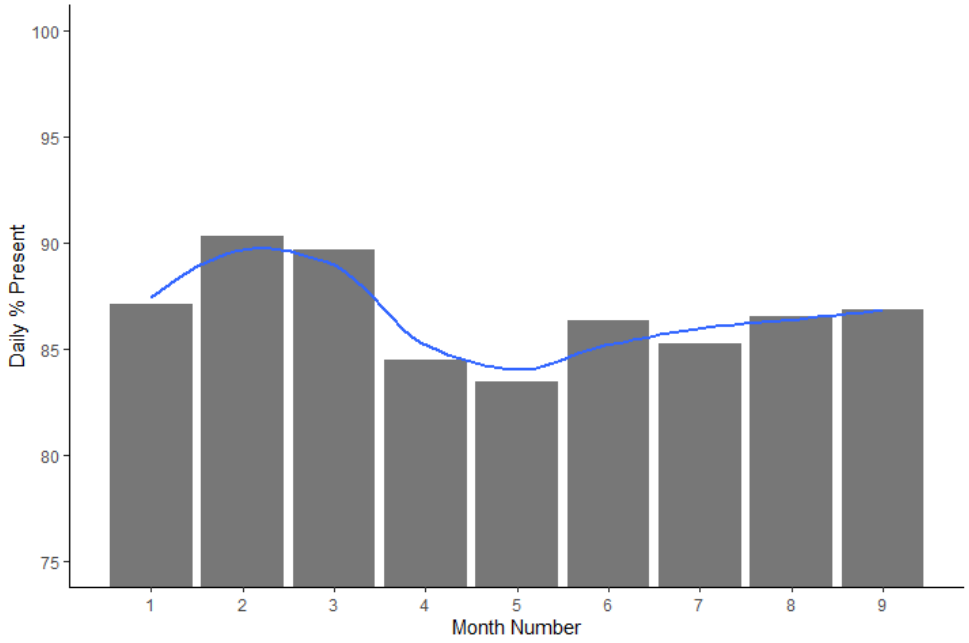

Figure 4.4. Example of seasonality in monthly attendance data. The blue line is a loess curve selected to highlight the seasonality.

After aggregating at the monthly scale there were 9-10 data points for each school depending on whether the school ended their year in May or June. I dropped the June data point as it was always a shortened month with uncharacteristic absenteeism. Similarly, I excluded the first week of school from the September data point for all schools because it also displayed uncharacteristic levels of absenteeism compared to the rest of the year. After these adjustments, all school-level data sets were composed of nine monthly data points.

Simulation of Attendance Effects

For all analyses, I randomly drew 3000 school-level data sets without replacement from the total collection of school-level monthly attendance data I assembled. Then, I simulated one of three different mean shifts (+1%, +5%, and +10%) in the monthly average attendance percentages. The first 3 points of data (September, October, November) were left unaltered to serve as calibration data. The mean shift was representative of an abrupt change in the underlying attendance generating process in early December. This simple effect was chosen based on prior studies of run charts and control charts in the literature (Anhøj, 2015; Anhøj & Wentzel-Larsen, 2018). It is likely that in a real-world improvement effort the effect would be more complicated. However, the aim of the present study is to show the potential of this method for capturing a simple effect in dynamic data, not to provide definitive evidence for the use of dynamic control charts with attendance data.

One difficulty in utilizing this approach is that simulating a mean shift in school-level monthly attendance data causes many schools to have over 100% attendance as the median school already has an average attendance rate of 92% (most kids in most schools are present

for most of the days). For example, a 10% increase in average attendance left 70% of all data points above 100. This is not ideal. One solution is to constrain the attendance to a maximum of 100% (even if the mean shift would put it over). However, this is a rather crude solution as it ignores the likely compression in attendance effects as the attendance percentage approaches 100%. Moreover, this would improperly mute the detection ability of the dynamic control chart because the effect on average would be smaller than a 10% increase. Another potential solution is to drop all schools from the analyses with greater than a 90% attendance rate. However, this is 74% of the sample and would leave too few school-year pairings (N = 2648) for simulation purposes. Furthermore, this would bias the results to schools with majority chronic absence where there are potentially different dynamics.

I decided not to make any adjustments to address this issue in the present study for two reasons. First, a whole school increase in average attendance of 5-10% is not a realistic effect because only chronically absent students can see that level of gain. However, for smaller within-school groupings this effect might be quite reasonable. For a single student, an increase of 10% is only two less absences per month (2 out of 20 days). Some chronically absent students are missing three or more days per month. This matters because 10% gains might be possible in select groups at many schools both realistically and without going over 100% average attendance within the select group. The second reason is the scope of this study. My aim is to show the potential of the dynamic control chart for substantiating improvement in data from dynamical processes. The dynamics of the data are present in all schools. The level of the attendance is of less importance. This issue, however, remains a limitation of this study.

Dynamic Control Charts

Dynamic control charts were produced using the R package dccharts (https://github.com/westdew/dccharts/). This package provides an R interface for using Bayesian structural time-series (Scott & Varian, 2014) to substantiate improvement in a manor akin to run chart and control chart usage. For each dynamic control chart, dccharts estimates a counterfactual for the post period based on the model chosen and the pre period data. Then, a p-value is calculated by examining where the actual data fell within the probability distribution of the estimated counterfactual. This Bayesian p-value indicates the proportion of the Bayesian simulations of the counterfactual where the data was more extreme than the actual post period data. It can be interpreted, similar to a frequentist p-value, as the probability of obtaining a result at least as extreme as the observed data, assuming the null hypothesis is true.

Recall, the dynamic control charts under analysis include: 1) a basic dynamic control chart with only a static intercept (i.e., the naive model), 2) a dynamic control chart with a local level component (i.e., a random walk), 3) a dynamic control chart including the school's previous year's attendance as a predictor, and 4) a dynamic control chart including the mean contemporaneous attendance of similar schools as a predictor. In the sections that follow, I explain each of these four types of dynamic control charts in greater detail before elaborating on the receiver operating characteristic (ROC) curve analysis used to compare the results.

Static Intercept

The counterfactual for the dynamic control chart with a static intercept (1) is estimated using a state-space model defined by the following equations:

$$y_t = \alpha_t + \epsilon_t$$
$$\alpha_t = M$$
$$\epsilon_t \sim N(0, \sigma_{\epsilon_t}^2)$$

The first equation is referred to as the observation equation. In this equation, y_t is the average attendance percentage for the t^{th} monthly observation, α_t is the underlying state of the attendance generation process, and ϵ_t is the independent normally distributed errors in the measurement of the state. The underlying state is defined in the second equation – known as the state equation – where, for a static intercept model, M is assumed to be constant. The parameters estimated by this model are the mean of the pre period, M, and the variance of the pre period, $\sigma_{\epsilon_t}^2$. All priors were specified as unassuming. Figure 4.5a presents an example of a dynamic control chart with a static intercept created by the package dccharts.

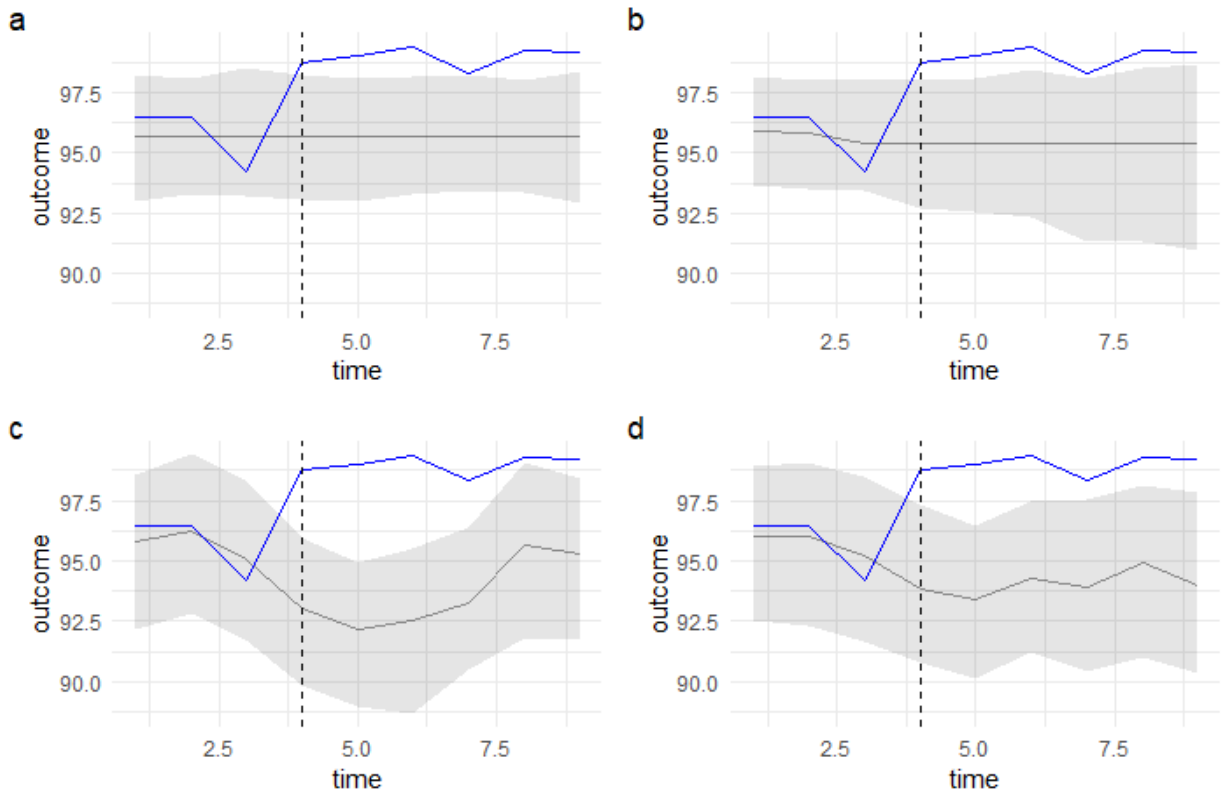

Figure 4.5. Examples of four types of dynamic control charts where non-random change was present and detected. The blue line graph is the monthly average attendance percentage data. The dashed line is the beginning of the non-random mean shift (+5). The grey line is the model fit to the left of the dashed line, and the counterfactual to the right of the dashed line. The grey ribbon is the 95% Bayesian credible interval. 4.5a. Dynamic control chart with a static intercept. 4.5b. Dynamic control chart with a local level component. 4.5c. Dynamic control chart including the school's previous year's attendance as a predictor. 4.5d. Dynamic control chart including the mean contemporaneous attendance of similar schools as a predictor.

Local Level

The counterfactual for the dynamic control chart with a local level (2) is estimated using a state-space model defined by the following equations:

$$y_t = \alpha_t + \epsilon_t$$
$$\alpha_t = \alpha_{t-1} + u_t$$
$$\epsilon_t \sim N(0, \sigma^2_{\epsilon_t})$$
$$u_t \sim N(0, \sigma^2_{u_t})$$

In the observation equation, y_t is the average attendance percentage for the t^{th} monthly observation, α_t is the underlying state of the attendance generation process, and ϵ_t is the independent normally distributed errors in the measurement of the state. In the state equation, α_{t-1} is the prior month's state, and u_t is a normally distributed stochastic innovation term. Put simply, this model assumes the level of the underlying state will float around similar to how it has in the past and penalizes prediction by widening the credible interval to include the dynamics. This penalty grows the further into the future one estimates. The parameters estimated by this model are the variance of the pre period, $\sigma_{\epsilon_t}^2$, and the variance of the innovations, $\sigma_{u_t}^2$. All priors were specified as unassuming. Figure 4.5b presents an example of a dynamic control chart with a local level created by the package dccharts.

Predictor

The counterfactual for the dynamic control charts with a covariate predictor (3, 4) is estimated using a state-space model defined by the following equations:

$$y_t = \alpha_t + \beta x_t + \epsilon_t$$
$$\alpha_t = M$$
$$\epsilon_t \sim N(0, \sigma_{\epsilon_t}^2)$$

In the observation equation, y_t is the average attendance percentage for the t^{th} monthly observation, α_t is the underlying state of the attendance generation process, β is the regression coefficient for the covariate predictor, x_t is the covariate predictor for the t^{th} monthly observation, and ϵ_t is the independent normally distributed errors in the measurement of the state. In the state equation, M is assumed to be constant if the predictor controls for most of the dynamics. Often a synthetic control model will include a local level

instead of a static intercept to adjust for any residual dynamicity. In supplemental analyses I found no benefits to this approach. The parameters estimated by this model are the regression coefficient for the covariate predictor, β, the mean of the pre period, M, and the variance of the pre period, $\sigma_{\epsilon_t}^2$. All priors were specified as unassuming.

For the dynamic control chart which includes the school's previous year's attendance as a predictor (3), x_t is a vector of the nine monthly attendance percentages from the prior calendar year for the school which produced the attendance data being tested. If schools have similar seasonality in attendance from year to year then this predictor should be a good basis for the counterfactual. Note, the 2012/13 data was specifically not used in creating the collection of school-level monthly attendance data sets so that all the years in the data would have a prior year for use in analyses. Figure 4.5c presents an example of this type of dynamic control chart created by the package dccharts.

For the dynamic control chart which includes the mean contemporaneous attendance of similar schools as a predictor (4), x_t is a vector of the average monthly attendance of the ten contemporaneous schools with the closest average daily attendance (ADA) to the school which produced the attendance data being tested. If schools have similar seasonality to other schools with comparable levels of attendance, then this predictor should be a good basis for the counterfactual. Figure 4.5d presents an example of this type of dynamic control chart created by the package dccharts.

Receiver Operating Characteristic (ROC) Curve Analysis

ROC curve analyses, from signal detection theory (Swets et al., 2000), will be used to assess the accuracy of the dynamic control charts. The ROC curve visually demonstrates the relationship between the sensitivity and the specificity of a diagnostic test. When performing a diagnostic test the goal is to determine if the data contains a signal or just noise. The sensitivity of the test indicates how well the diagnostic detects signals (e.g., a 0.80 sensitivity would mean 80% of the signals were detected correctly and there was a 20% false negative rate). A diagnostic test can be tuned to be more sensitive. However, this typically has a trade off with the test's specificity – the ability of the test to detect noise correctly (e.g., a 0.95 specificity would mean 95% of the noise was detected correctly and there was a 5% false positive rate). A helpful question to understand this relationship better is to consider how a diagnostic test might easily become maximally sensitive. The absurd answer, assume everything is a signal. This would result in a sensitivity of 1 as every signal would be correctly detected. However, the specificity, would now be very poor as all the noise was incorrectly specified as a signal.

By tuning the diagnostic threshold across the full domain of possible cuts many different sensitivity-specificity pairings can be generated. A ROC curve is the plot of these pairings. Typically, $sensivity$ is plotted on the y-axis and $(1 - specificity)$ is plotted on the x-axis. For an example, see Figure 4.6. An ROC curve above the diagonal indicates a predictive diagnostic. The more of the ROC space the curve encompasses the higher the test's accuracy. If the test is no better than a coin flip, then the ROC curve will simply overlap the diagonal, obscuring half the area of the ROC space. Finally, an ROC curve below the diagonal indicates a diagnostic which predicts the opposite of what it claims.

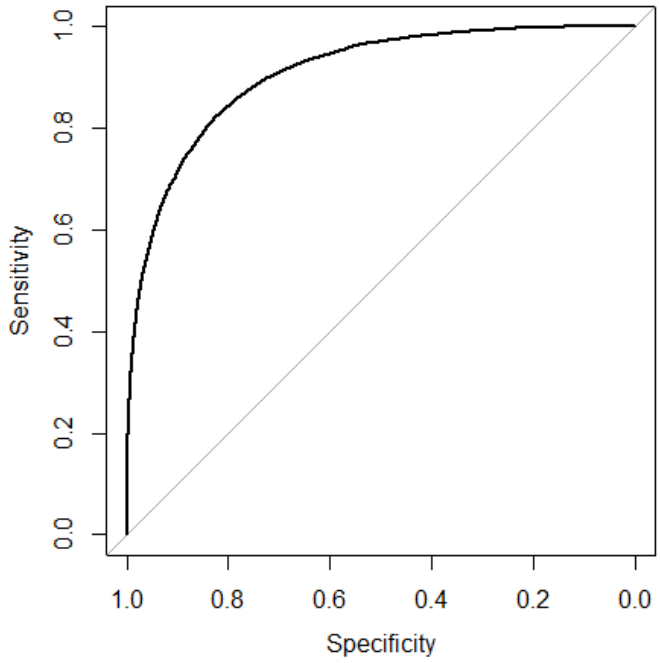

Figure 4.6. Example of a symmetrical ROC curve with an AUC of 0.90.

A useful statistic which can be calculated from the ROC curve is the area under the curve (AUC). The AUC can be interpreted as the probability of correctly identifying the signal when presented with two data sets, one with a signal and one with only noise (Hanley & McNeil, 1982). For a perfect diagnostic test which fills the ROC space the area will be 1, indicating the probability is 100%. However, for an ineffective diagnostic test which overlaps the diagonal the area will be 0.5, indicating a 50-50 chance.

The AUC of the ROC curves will be evaluated in two ways. First, the AUC will be compared to benchmarks from the educational literature. In the response to intervention field (RTI), the standard set by the National Center on Response to Intervention (2010) indicates that an AUC above 0.85 is considered “convincing evidence” that a screening test accurately classifies, while an AUC from 0.75 to 0.85 is considered “partially convincing evidence,” and an

AUC below 0.75 is “unconvincing evidence.” The field of curriculum-based measurement (CBM) sets a slightly higher standard. The criteria for an excellent CBM is an AUC between 0.90 and 1.0, while 0.80 to 0.89 is a good CBM, and 0.70 to 0.79 is a poor CBM (Christ et al., 2013). For the present study, I decided to use the higher standard from the CBM literature, as CBMs are arguably similar to continuous improvement’s practical measures (Yeager et al., 2013).

The second way the AUC statistics will be evaluated is in comparing dynamic control charts to each other. To determine if two AUC statistics, AUC_1 and AUC_2, are significantly different a z-score can be calculated using the following formula:

$$z = \frac{AUC_1 - AUC_2}{s}$$

where s is the standard deviation of the differences between the AUC statistics upon repeated sampling. This standard deviation can be estimated using a bootstrapping technique (Robin et al., 2011). Once the z score is calculated it can be compared to a normal distribution to determine whether the AUC statistics are significantly different. All ROC analyses (ROC curve plots, AUC statistics, and AUC significance tests) were computed using the pROC package (Robin et al., 2011) in the R programming language (R Core Team, 2020).

Results

Recall, the aim of the present study is to show the potential of the dynamic control chart for substantiating improvement in data from dynamical processes. To this end, four types of dynamic control charts utilizing different tools for accommodating dynamicity were tested on real-world attendance data with simulated mean shift effects. The ROC analyses of the four

types of dynamic control charts for the three effects (+1%, +5%, and +10%) produced 12 ROC curves. These ROC curves and their AUC statistics are examined for each dynamic control chart type in the sections that follow.

Dynamic control chart with a static intercept (1)

Figure 4.7 presents the ROC curve analysis of the static intercept model. This naive model did not fit the attendance data well. For the smallest effect, +1%, the AUC was just 0.291, indicating the test predicted the opposite of what was intended. This most likely occurred because the effect was overwhelmed by opposing dynamics, e.g., declining attendance during the year. Figure 4.8a presents an example where this phenomenon is clearly visible. For the largest effect, +10%, the AUC, 0.695, fell short of even the benchmark for a poor measure, 0.7. The static intercept model, which is similar to a run chart or control chart, appears to be an impoverished diagnostic test for substantiating improvement in monthly attendance data.

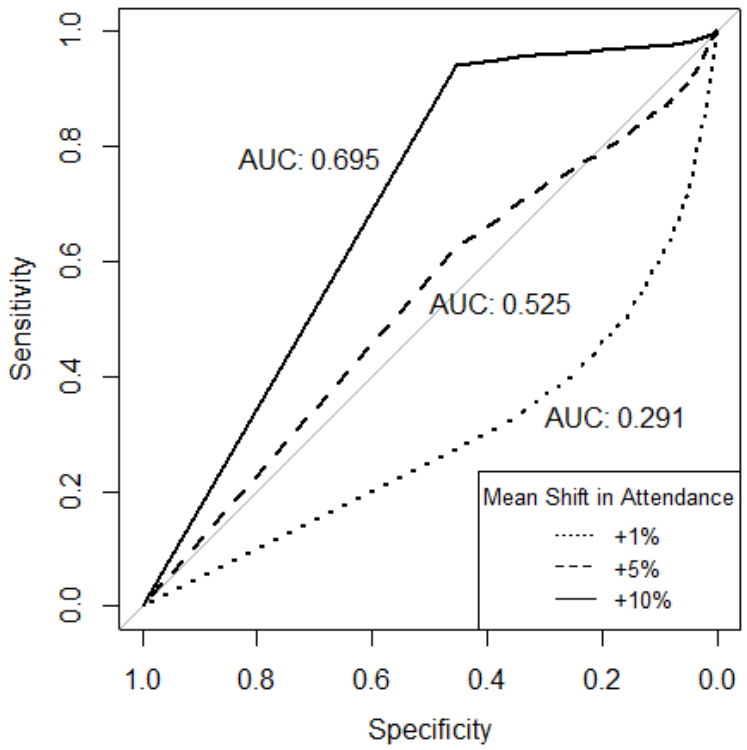

Figure 4.7. ROC curve of dynamic control chart with a static intercept for various mean shift effects (+1%, +5%, and +10%).

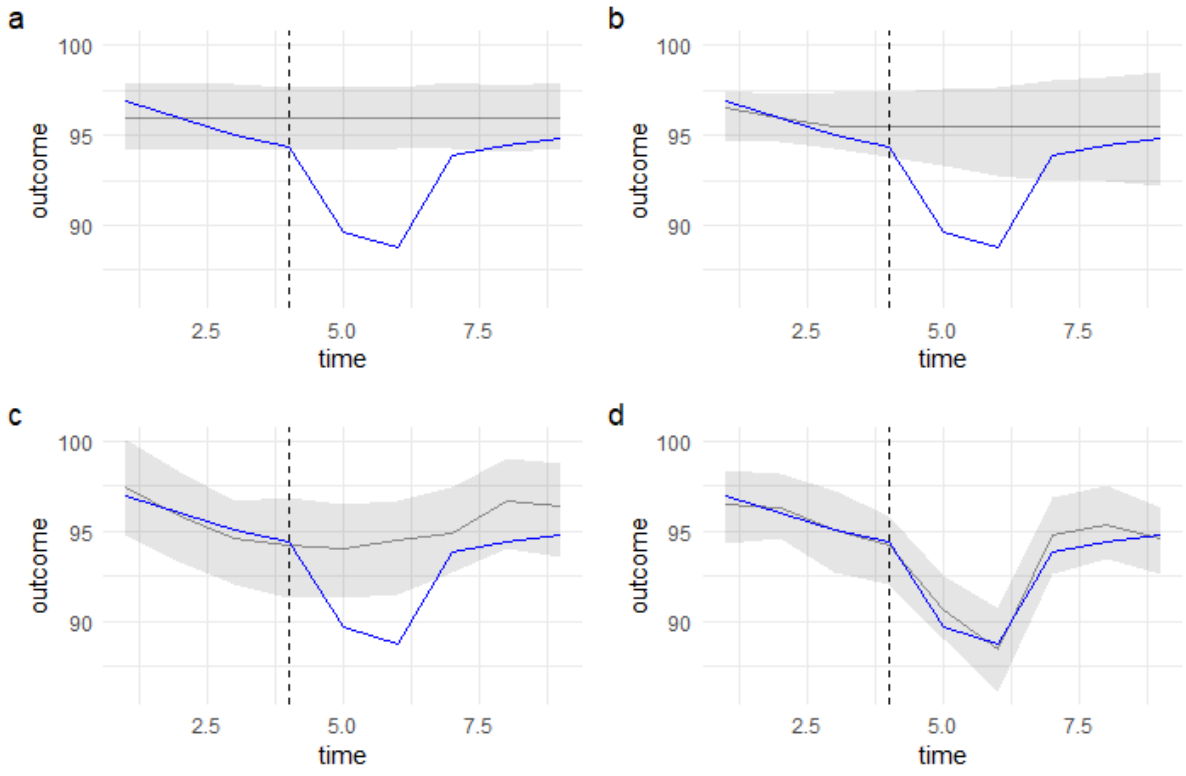

Figure 4.8. Examples of four types of dynamic control charts applied to unaltered data with dynamicity. The blue line graph is the monthly average attendance percentage data. The dashed line is the end of the calibration. The grey line is the model fit to the left of the dashed line, and the counterfactual to the right of the dashed line. The grey ribbon is the 95% Bayesian credible interval. 4.8a. Dynamic control chart with a static intercept. 4.8b. Dynamic control chart with a local level component. 4.8c. Dynamic control chart including the school's previous year's attendance as a predictor. 4.8d. Dynamic control chart including the mean contemporaneous attendance of similar schools as a predictor.

Dynamic control chart with a local level component (2)

Figure 4.9 presents the ROC curve analysis of the local level model. This model performed marginally better than the static intercept model. The AUC of the smallest effect was still below 0.5 indicating the diagnostic was reversed. As can be seen in Figure 4.8b, the penalized credible interval still misses most of the underlying dynamics, even though it now extends down towards the mid-year decline in attendance. For larger effects, this model

improved. The AUC of the largest effect, 0.868, was substantively and significantly greater than the AUC of the intercept model, $p < 0.001$. Moreover, for this effect the dynamic control chart with a local level component would be considered a good measure for a CBM.

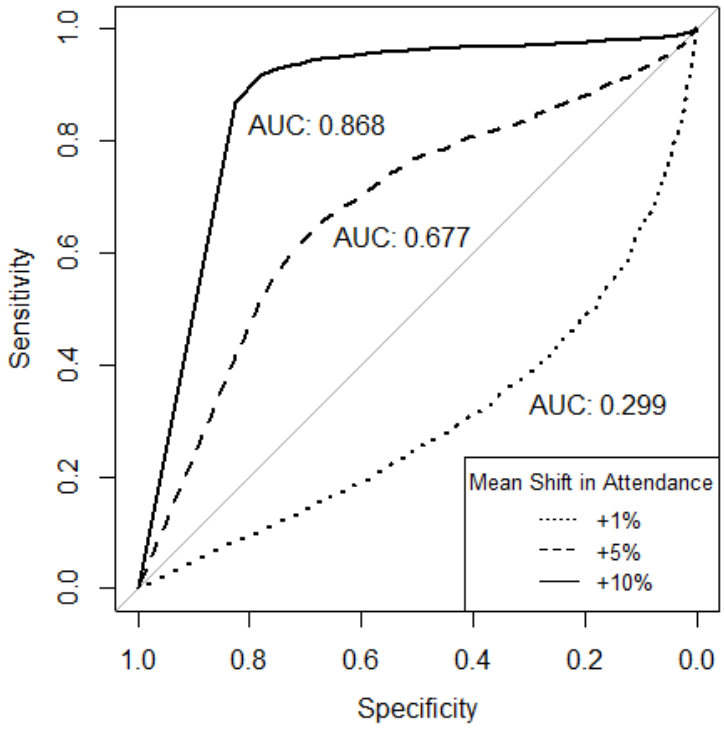

Figure 4.9. ROC curve of dynamic control chart with a local level component for various mean shift effects (+1%, +5%, and +10%).

Dynamic control chart including the school's previous year's attendance as a predictor (3)

Figure 4.10 presents the ROC curve analysis of the prior year predictor model. This model performed much better than the previous two. For the smallest effect, the model appears to fit the attendance data, resulting in an AUC which was slightly better than chance, 0.6. Though, as Figure 4.8c shows, the prior year was not able to predict all the dynamics of the present year. The AUC statistics for the moderate, +5%, and large, +10%, effects were 0.91 and

0.969, respectively. Both were excellent measures by the standards of the CBM literature, and both were significant improvements over all prior models, $p < 0.001$.

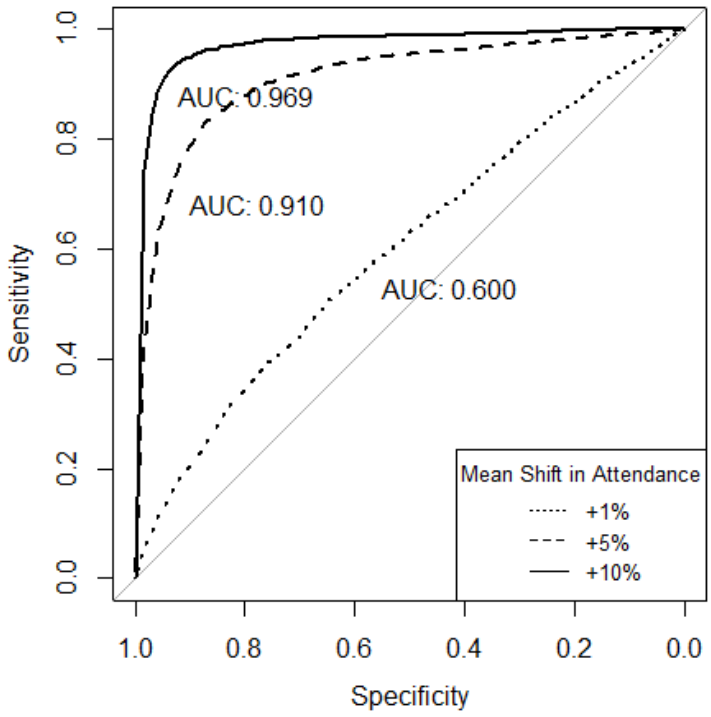

Figure 4.10. ROC curve of a dynamic control chart including the school's previous year's attendance as a predictor for various mean shift effects (+1%, +5%, and +10%).

Dynamic control chart including the mean contemporaneous attendance of similar schools as a predictor (4)

Figure 4.11 presents the ROC curve analysis of the contemporaneous school predictor model. This model appears to have superior accuracy compared to the other dynamic control charts. Figure 4.8d presents an example of how well this model could fit the data and predict the counterfactual in some cases. The AUC statistics for the moderate, +5%, and large, +10%, effects were 0.978 and 0.995, respectively. These AUC statistics indicate the dynamic control chart was an excellent measure. Both AUC statistics were significant improvements over all

prior models, $p < 0.001$. Furthermore, the improvement was large enough that the accuracy of the contemporaneous school predictor model for a moderate effect, +5%, was statistically indistinguishable from the accuracy of the prior year predictor model for a large effect, +10%, $p = 0.282$.

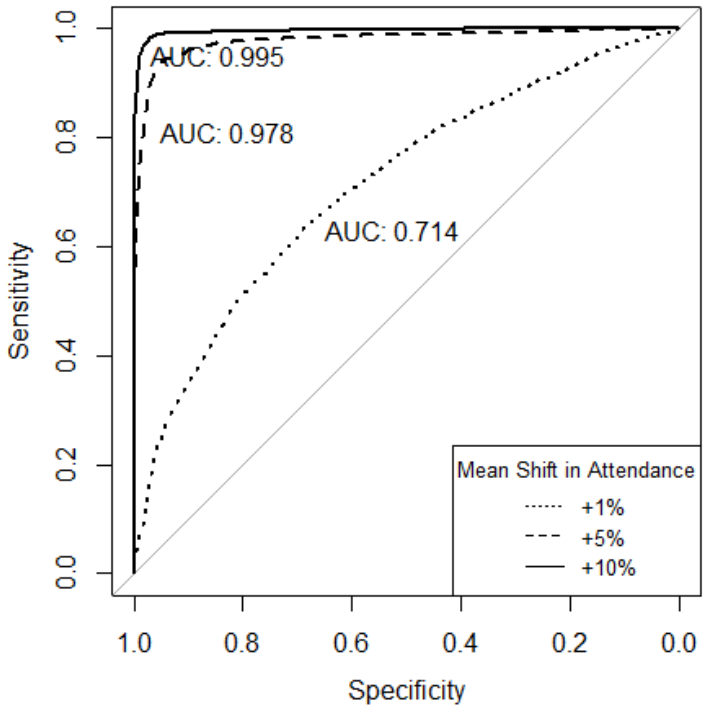

Figure 4.11. ROC curve of dynamic control chart including the mean contemporaneous attendance of similar schools as a predictor for various mean shift effects (+1%, +5%, and +10%).

Discussion

The purpose of the present study was to show the potential of the dynamic control chart for substantiating improvement in data from dynamical processes. Prior investigations (see Chapter 3) had raised questions around the ability of dynamic control charts to detect realistic effects in dynamic data. In this study, four types of dynamic control charts were

examined: 1) a basic dynamic control chart with only a static intercept (i.e., the naive model), 2) a dynamic control chart with a local level component (i.e., a random walk), 3) a dynamic control chart including the school's previous year's attendance as a predictor, and 4) a dynamic control chart including the mean contemporaneous attendance of similar schools as a predictor. The static intercept and local level models are similar to the models which performed poorly with dynamic data in prior investigations (see Chapter 3). The predictor-based models, however, had not been studied before. These models used covariate predictors to control for the seasonal variance in attendance.

The static intercept model and the local level model were found to be insufficient. Even for an increase of 10% in average attendance, the accuracy of these models was below rigorous standards. The predictor-based models, though, proved to be excellent diagnostic tests. For an increase of 5% or more in average attendance, the area under the curve (AUC) of the ROC analyses were all larger than 0.9, the standard for an excellent measure in curriculum-based measurement (CBM). Though a 5% increase in school-level attendance may be difficult to imagine, this level of change is just one less day absent per month for every student. This is possible, especially for smaller more targeted groups of students. Many students in a school will not need a 5% change in their attendance. Increases in aggregate attendance will be at least 5% if the aggregate is from a targeted group of chronically absent students and the change ideas are effective.

For a 5% increase in average attendance, the best diagnostic test in the present study was the dynamic control chart including the mean contemporaneous attendance of similar schools as a predictor, AUC = 0.978. This control chart was significantly better than all the

alternatives, $p < 0.001$. Figure 4.12 presents this case visually, comparing the ROC curves of the four types of dynamic control charts. This is an interesting finding as it suggests that the dynamics in attendance during a school year are more closely related to the events in the district that year than to returning school specific seasonal effects. Most likely, the truth is a compromise between the two. Future work should consider ensemble dynamic control charts which include both historical and contemporaneous covariates, potentially allowing their inclusion to be determined empirically by the data (George & McCulloch, 1997).

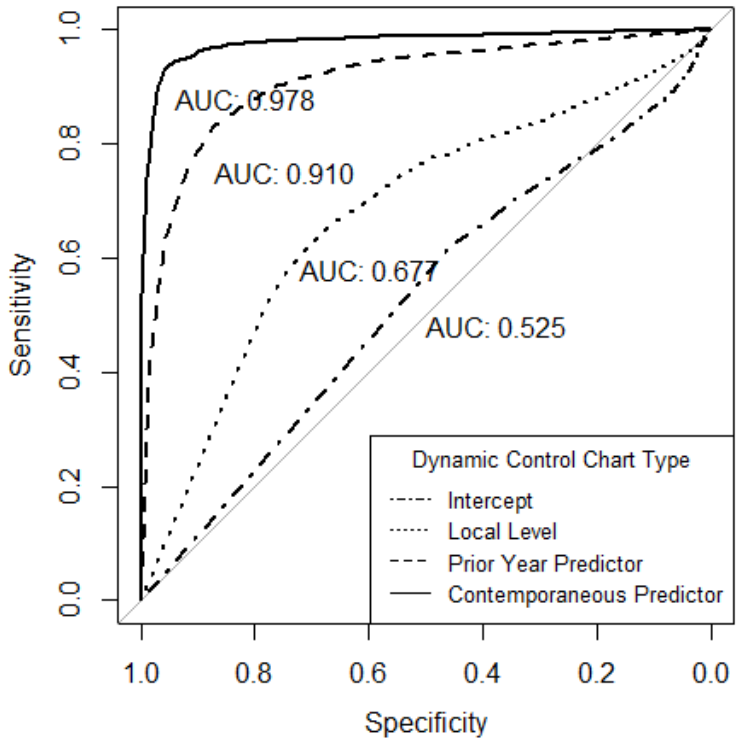

Figure 4.12. ROC curves of four types of dynamic control charts for a mean shift of +5%.

There were some clear limitations with the present study. Simulations are useful for comparing methods under controlled conditions. However, often in the field unexpected and complicated conditions are encountered. This study did much to mitigate these concerns by

using authentic school-level attendance data. Though, this study also assumed an overly simplistic mean shift effect and allowed the attendance to rise above 100%. These decisions raise questions around the actual performance of the dynamic control chart in an attendance improvement context. Future research should consider examining more gradual effects and using student-level data to calculate the average monthly attendance of only chronically absent students. Even with these adjustments this study would not be able to say for certain that the diagnostic values and relationships observed in these simulations would replicate using data from real world scenarios. The strength of this method in the field is best confirmed through its effective use in practice.

To that point, the evidence presented in the present study suggests that continuous improvement professionals should consider using the Dynamic Control Chart when attempting to substantiate improvement in data from dynamical processes. The dynamic control chart is a largely automated method for establishing evidence of improvement which requires little statistical skill, minimal data collection before and after a change, and, as was seen in the present study, has the tools to accommodate different forms of temporal dynamicity.

Chapter 5 – Afterword

This dissertation presented three academic studies (Chapters 2-4) focused on the crucial task of substantiating improvement in educational continuous improvement work. Recall, the ability to quickly provide strong evidence that a change led to improved outcomes is vital to the success of continuous improvement methods (Reed & Card, 2016).

In Chapter 2, I addressed the challenge of translating continuous improvement methods from industry to educational contexts. To do so I asked three questions: 1) What are the current methods used for substantiating improvement in industry and health care? 2) What are the differences between education and sectors like industry and health care that affect the feasibility of substantiating improvement using existing methods? 3) What characteristics are necessary in a method for substantiating improvement in education?

In addressing the first question, I discovered that statistical process control methods have, in practice, not changed much in the last century despite advancements in the methodological literature and shifts to new contexts like modern industrial applications and health care. I also found that the statistical process control methods used in industry have many strengths which will be needed in educational improvement work including the ability to function with small purposeful samples, generate causal evidence in experimental contexts, and provide practical feedback to users. I then examined how people, organizations, data, and research differ between education and other sectors like industry and health care. I found many impactful differences. The most important for substantiating improvement being the slow cadence of educational data, the dynamicity in educational processes, and the limited

training of educators in statistics and research. Finally, I identified five requirements of a method for substantiating improvement in education based on the existing strengths of statistical process control and the relevant differences between industry and education. Specifically, I found this method needed to be disciplined but pragmatic, appropriate for small samples, responsive with limited data, semiparametric, and unobtrusive and automated. After a consideration of current methods in education and industry, I concluded that most methods were inadequate, and a new method would be required to meet this list of often contradictory methodological needs.

In Chapter 3, I introduced the Dynamic Control Chart (https://github.com/westdew/dccharts/) – a diagnostic test for substantiating improvement (made for this dissertation) that is based on statistical techniques designed for making causal claims using automated short-term forecasts (Brodersen et al., 2015). I argued that the dynamic control chart was novel in that it was fully automated and could be easily adjusted to accommodate dynamics in the data. Moreover, I suggested the dynamic control chart was a much stronger test of change than the decision rules used in statistical process control methods. To confirm this assertion, I conducted a simulation study. My results confidently supported my claim. For normally distributed exchangeable data and weakly autocorrelated data, the base dynamic control chart performed substantially better than run charts and control charts under all tested circumstances. In this chapter, I also examined the performance of the autoregressive component for dynamic control charts. Unfortunately, and somewhat unexpectedly, the autoregressive component underperformed when detecting a large mean shift in simulations of moderate to heavily autocorrelated data.

Finally, in Chapter 4, I revisited my claim that dynamic control charts can be easily adjusted to accommodate dynamics in the data. Given the underwhelming performance of the dynamic control chart with an autoregressive component this statement required further investigation. Again, I conducted a simulation study. However, this time I simulated effects in real-world school-level attendance data featuring autocorrelation, nonstationarity, and fractality (Koopmans, 2016). By using real data, I could investigate both the nonparametric and the semiparametric tools available for modeling dynamical processes. My results in Chapter 4 supported the claim that dynamic control charts can be easily adjusted to accommodate dynamics. In particular, for the attendance data used in the present study, models with a historical or contemporaneous predictor to adjust for the dynamics proved to be excellent diagnostic tests for substantiating improvement in monthly average attendance data.

The present dissertation demonstrated that the Dynamic Control Chart has the potential to serve the same role as existing SPC methods in substantiating change in continuous improvement work. Moreover, the simulations presented suggest that the Dynamic Control Chart could be a superior diagnostic test in many circumstances. Professionals working in continuous improvement should consider introducing the Dynamic Control Chart into their practice. Run charts and control charts are widely employed in improvement because they are straightforward and easy to use. Dynamic Control Charts have the promise to fill the same role while also providing stronger confirmatory tests, supporting any length of pre or post data, and modelling dynamic data.

Conclusion

Public schools in the United States have been subject to reform efforts since their inception in the 19th century. Some of these efforts have changed education for the better (Cohen & Mehta, 2017), but many have had little effect even as they were attempted again and again (Cuban, 1990). Today, continuous improvement is one reform among many vying for a scarce pool of educational resources (made available by the recession of testing and accountability from the reform limelight). Only time will tell which of these reforms will benefit students the most. However, at this moment, after a decade of growth in the use of continuous improvement in schools and a steady increase in political, academic, and monetary support for continuous improvement in the educational sector, continuous improvement is well positioned to become a key educational initiative for years to come if it can show some evidence of success. Furthermore, continuous improvement could finally realize a decades long shift in educational reform towards an emphasis on local knowledge and actors. This opportunity should not be squandered. The question, then, for proponents of continuous improvement is: What is required for continuous improvement to succeed?

I argued in the present dissertation that the translation of continuous improvement to education from industry lost a crucial aspect of the method, namely a way to substantiate that a change was an improvement. Industry had a clear set of statistical tools for gaining control of a process and then monitoring the process. This allowed engineers to determine whether a change made in a process improved the process. Continuous improvement in education has retained the importance of this step in disciplined improvement work, but the tools used in industry have not made it to education, in part, because they were unworkable for educators

using educational data. The run chart is the only tool of this nature discussed in the educational literature.

Continuous improvement will not work effectively without a disciplined, scientific method for substantiating that a change is an improvement. All the continuous improvement methods in the educational literature utilize an iterative process to solve problems (Yurkofsky et al., 2020). The success of iteration in building improvement over time is contingent on the quality of the feedback into the next cycle (Reed & Card, 2016). Put simply, educators must determine whether the changes they make are improvements in order to decide whether to adopt or abandon those changes going forward. Without a clear method to assist in making these decisions, the ability of an educator to improve becomes contingent on their idiosyncratic talent for determining improvement. At best this will merely slow the rate at which education improves. However, eventually, this will also limit how much education can improve, once adjustments on average stop progressing practice (i.e., one step forward and one step back). Furthermore, the results of continuous improvement will likely end up as heterogenous as measures like teacher valued-added.

Presently, continuous improvement encompasses a large set of new ideas and practices in education (Yurkofsky et al., 2020) which could benefit directly from Dynamic Control Charts. However, continuous improvement also likely includes many long-standing educational practices like formative assessment and reflective practice (Schön, 1992), and fundamentally, continuous improvement is embedded in the concept of education itself, as an educator's aim is to improve their students in some manner. Furthermore, continuous improvement occurs on many different levels of work in education (e.g., district-wide initiatives, collaborative

professional communities, solo teachers). Consequently, even in day-to-day work, educators are often left wondering whether the changes they make are improvements. Answering this question is a real problem for educators which continuous improvement makes manifest.

The Dynamic Control Chart and the research presented in the present dissertation are the beginnings of a solution to this problem. I have put forth a statistical tool – the dccharts R library (https://github.com/westdew/dccharts/) – which implements the Dynamic Control Chart and can be used to substantiate improvement in education at many different levels. I have argued that the Dynamic Control Chart meets the methodological requirements I identified for such a method, and I presented simulations that support my arguments. Specifically, I found in simulations that the Dynamic Control Chart was better suited than traditional alternatives to handle limited data (see Chapter 3) and dynamic data (see Chapter 4) – two common occurrences in education.

For the Dynamic Control Chart to be used more broadly there are some immediate problems that would need to be addressed. First, although the method has the potential for automation, it is currently accessible only through the statistical programming language R, and educators are not trained in statistics or programming. The Dynamic Control Chart will need a different interface that is amiable to educators but retains the benefits of the R package. Second, many of the more powerful techniques for substantiating change in dynamic data (see Chapter 4) will require the use of prior year comparisons and contemporaneous comparisons. Educators do not currently have access to the data they would need to generate these comparisons. Districts will need to develop a process for educators to utilize these data streams to take full advantage of a method like the Dynamic Control Chart. This will be a significant

challenge of the work, especially in the lower levels of continuous improvement (e.g., the lone teacher) as sharing raw student data has many practical and ethical concerns. Finally, educators will need training in continuous improvement and in the use of the Dynamic Control Chart in supporting continuous improvement. Though, the right interface could reduce the need for training dramatically, especially if the method is couched in terms of an existing common practice like formative assessment.

Continuous improvement has the potential to change education for the better. Evidence-based reform has been attempting to increase the rigor of research on educational programs for decades to spark a scientific revolution but has struggled to bring research into practice. Continuous improvement, by drawing on local actors and knowledge, will bring scientific rigor directly to educators' work. Educators make necessary decisions every day as a part of their jobs. Continuous improvement ensures that these decisions made at every level of education – whether they are small instructional changes or the adoption of proven programs district wide – lead towards improvement for students. Widespread use of continuous improvement in schools is the type of paradigm shift which might truly lead to a scientific revolution in education. However, the cornerstone of continuous improvement work is the ability to determine if a change was an improvement, and, presently, answering this question is a real problem for educators, even in the district office. The Dynamic Control Chart is the beginnings of a solution to this pressing problem.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, *59*(2), 495-510. https://doi.org/10.1111/ajps.12116

Anhøj, J. (2015). Diagnostic value of run chart analysis: using likelihood ratios to compare run chart rules on simulated data series. *PLoS One*, *10*(3), e0121349. https://doi.org/10.1371/journal.pone.0121349

Anhøj, J. (2018). qicharts2: Quality improvement charts for r. *Journal of Open Source Software*, *3*(25), 699. https://doi.org/10.21105/joss.00699

Anhøj, J., & Wentzel-Larsen, T. (2018). Sense and sensibility: on the diagnostic value of control chart rules for detection of shifts in time series data. *BMC medical research methodology*, *18*(1), 1-8. https://doi.org/10.1186/s12874-018-0564-0

Anjard, R. P. (1995). SPC chart selection process. *Microelectronics Reliability*, *35*(11), 1445-1447. https://doi.org/10.1016/0026-2714(95)00119-M

Attia, J. (2003). Moving beyond sensitivity and specificity: using likelihood ratios to help interpret diagnostic tests. *Australian Prescriber*, *26*(5), 111-113. https://doi.org/10.18773/austprescr.2003.082

Balfanz, R., & Byrnes, V. (2012). Chronic absenteeism: Summarizing what we know from nationally available data. *Baltimore: Johns Hopkins University Center for Social Organization of Schools*, *1*(1), 1-46.

Barnard, G. A. (1959). Control charts and stochastic processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, *21*(2), 239-257. https://doi.org/10.1111/j.2517-6161.1959.tb00336.x

Berliner, D. (2013). Effects of inequality and poverty vs. teachers and schooling on America's youth. *Teachers College Record*, *115*(12), 1-26.

Biesta, G. (2009). Good education in an age of measurement: On the need to reconnect with the question of purpose in education. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)*, *21*(1), 33-46. https://doi.org/10.1007/s11092-008-9064-9

Biesta, G. J. (2010). Why 'what works' still won't work: From evidence-based education to value-based education. *Studies in philosophy and education*, *29*(5), 491-503.

https://doi.org/10.1007/s11217-010-9191-x

Borman, G. D., Grigg, J., Rozek, C. S., Hanselman, P., & Dewey, N. A. (2018). Self-affirmation effects are produced by school context, student engagement with the intervention, and time: Lessons from a district-wide implementation. *Psychological Science*, *29*(11), 1773-1784. https://doi.org/10.1177/0956797618784016

Bowers, A. J., & Zhou, X. (2019). Receiver operating characteristic (ROC) area under the curve (AUC): A diagnostic measure for evaluating the accuracy of predictors of education outcomes. *Journal of Education for Students Placed at Risk (JESPAR)*, *24*(1), 20-46. https://doi.org/10.1080/10824669.2018.1523734

Box, G., & Narasimhan, S. (2010). Rethinking statistics for quality control. *Quality Engineering*, *22*(2), 60-72. https://doi.org/10.1080/08982110903510297

Bravata, D. M., Gienger, A. L., Holty, J. E. C., Sundaram, V., Khazeni, N., Wise, P. H., ... & Owens, D. K. (2009). Quality improvement strategies for children with asthma: a systematic review. *Archives of pediatrics & adolescent medicine*, *163*(6), 572-581. https://doi.org/10.1001/archpediatrics.2009.63

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. *Annals of Applied Statistics*, *9*(1), 247-274.

https://doi.org/10.1214/14-AOAS788

Bryk, A. S., & Gomez, L. M. (2008). Ruminations on reinventing an R&D capacity for educational improvement. *The future of educational entrepreneurship: Possibilities of school reform*, 181-206.

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Harvard Education Press.

Carey, R. G. (2002). How do you know that your care is improving? Part I: basic concepts in statistical thinking. *The Journal of Ambulatory Care Management*, *25*(1), 80-87.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological bulletin*, *132*(3), 354. https://doi.org/10.1037/0033-2909.132.3.354

Champ, C. W., & Woodall, W. H. (1987). Exact results for Shewhart control charts with supplementary runs rules. *Technometrics*, *29*(4), 393-399.

Chen, Q., Kruger, U., & Leung, A. Y. (2009). Cointegration testing method for monitoring nonstationary processes. *Industrial & Engineering Chemistry Research*, *48*(7), 3533-3543.

https://doi.org/10.1021/ie801611s

Chen, Z. (2010). A note on the runs test. *Model Assisted Statistics and Applications*, *5*(2), 73-77. https://doi.org/10.3233/MAS-2010-0142

Chi, M. T., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive science*, *25*(4), 471-533. https://doi.org/10.1207/s15516709cog2504_1

Chirume, E. (2018). Instructional Leaders and Understanding Data: The Status and Prevalence of Research and Statistics Courses in a Midwestern State's Educator Preparation Programs. *American Journal of Educational Research*, *6*(7), 997-1004. https://doi.org/10.12691/education-6-7-16

Christ, T. J., Zopluoglu, C., Monaghen, B. D., & Van Norman, E. R. (2013). Curriculum-based measurement of oral reading: Multi-study evaluation of schedule, duration, and dataset quality on progress monitoring outcomes. *Journal of School Psychology*, *51*(1), 19-57. https://doi.org/10.1016/j.jsp.2012.11.001

Chubb, J. E., & Moe, T. M. (1991). Politics, markets and America's schools.

Coburn, C. E., Touré, J., & Yamashita, M. (2009). Evidence, interpretation, and persuasion: Instructional decision making at the district central office. *Teachers College*

Record, *111*(4), 1115-1161.

Cochran-Smith, M. (2005). Studying Teacher Education: What We Know and Need to Know. *Journal of Teacher Education*, *56*(4), 301–306. https://doi.org/10.1177/0022487105280116

Cohen, D. K., & Mehta, J. D. (2017). Why Reform Sometimes Succeeds: Understanding the Conditions That Produce Reforms That Last. *American Educational Research Journal*, *54*(4), 644–690. https://doi.org/10.3102/0002831217700078

Cohen, G. L., Garcia, J., & Goyer, J. P. (2017). Turning point: Targeted, tailored, and timely psychological intervention. In A. J. Elliot, C. S. Dweck, & D. S. Yeager (Eds.), *Handbook of Competence and Motivation, Second Edition*. Guilford Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cohen-Vogel, L., Tichnor-Wagner, A., Allen, D., Harrison, C., Kainz, K., Socol, A. R., & Wang, Q. (2014). Implementing educational innovations at scale: Transforming researchers into continuous improvement scientists. *Educational Policy*, *29*(1), 257-277. https://doi.org/10.1177/0895904814560886

Cuban, L. (1990). Reforming Again, Again, and Again. Educational Researcher, 19(1), 3–13. https://doi.org/10.3102/0013189X019001003

D'Agostino, J. V., Rodgers, E., & Mauck, S. (2018). Addressing inadequacies of the observation survey of early literacy achievement. *Reading Research Quarterly*, *53*(1), 51-69. https://doi.org/10.1002/rrq.181

De Ketelaere, B., Mertens, K., Mathijs, F., Diaz, D. S., & Baerdemaeker, J. D. (2011). Nonstationarity in statistical process control—issues, cases, ideas. *Applied Stochastic Models in Business and Industry*, *27*(4), 367-376. https://doi.org/10.1002/asmb.911

De Ketelaere, B., Rato, T., Schmitt, E., & Hubert, M. (2016). Statistical process monitoring of time-dependent data. *Quality Engineering*, *28*(1), 127-142. https://doi.org/10.1080/08982112.2015.1100474

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2-21. https://doi.org/10.1016/j.socscimed.2017.12.005

Deming, W. E. (2000). *The new economics: For industry, government, education* (2nd ed.). MIT Press.

Dewey, J. (1910). *How we think*. D.C. Heath & Company.

DiMaggio, P., & Powell, W. W. (1983). The iron cage revisited: Collective rationality and institutional isomorphism in organizational fields. *American sociological review*, *48*(2), 147-160. https://doi.org/10.2307/2095101

DuPaul, G. J., Eckert, T. L., & Vilardo, B. (2012). The effects of school-based interventions for attention deficit hyperactivity disorder: A meta-analysis 1996–2010. *School psychology review*, *41*(4), 387-412. https://doi.org/10.1080/02796015.2012.12087496

Durbin, J., & Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford university press.

Eddington, A. S. (1929). *The nature of the physical world*. Cambridge, England: Cambridge University Press.

Elmore, R. F. (2016). "Getting to scale…" it seemed like a good idea at the time. *Journal of Educational Change*, *17*(4), 529-537. https://doi.org/10.1007/s10833-016-9290-8

Ferrer, A. (2014). Latent structures-based multivariate statistical process control: A paradigm shift. *Quality Engineering*, *26*(1), 72-91. https://doi.org/10.1080/08982112.2013.846093

Fisher, R. A. (1937). *The design of experiments*. Edinburgh, England: Oliver and Boyd.

García, E., & Weiss, E. (2018). Student Absenteeism: Who Misses School and How Missing School Matters for Performance. *Economic Policy Institute*.

George, E. I., & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, *7*(2), 339–373.

Gibbons, M. T. (2020). Higher Education R&D Funding from All Sources Increased for the Third Straight Year in FY 2018. Washington, DC: National Science Foundation.

Gierut, J. A., Morrisette, M. L., & Dickinson, S. L. (2015). Effect size for single-subject design in phonological treatment. *Journal of Speech, Language, and Hearing Research*, *58*(5), 1464-1481. https://doi.org/10.1044/2015_JSLHR-S-14-0299

Grunow, A., Hough, H., Park, S., Willis, J., & Krausen, K. (2018). Towards a Common Vision of Continuous Improvement for California. Technical Report. Getting Down to Facts II. *Policy Analysis for California Education, PACE*.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.

https://doi.org/10.1148/radiology.143.1.7063747

Hanselman, P., Rozek, C. S., Grigg, J., & Borman, G. D. (2017). New evidence on self-affirmation effects and theorized sources of heterogeneity from large-scale replications. *Journal of educational psychology*, *109*(3), 405. https://doi.org/10.1037/edu0000141

Hess, F. M. (2011). *Spinning Wheels: The Politics of Urban School Reform*. Brookings Institution Press.

Honig, M. I. (Ed.). (2006). *New directions in education policy implementation: Confronting complexity*. Suny Press.

Hough, H., Willis, J., Grunow, A., Krausen, K., Kwon, S., Mulfinger, L. S., & Park, S. (2017). Continuous Improvement in Practice. *Policy Analysis for California Education, PACE*.

Imbens, G. W., & Kolesar, M. (2016). Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics*, *98*(4), 701-712. https://doi.org/10.1162/REST_a_00552

Ingersoll, R. M., & Perda, D. (2008). The status of teaching as a profession. *Schools and society: A sociological approach to education*, 106-118.

Koopmans, M. (2015). When Time Makes a Difference: Addressing Ergodicity and Complexity in Education. *Complicity: An International Journal of Complexity and Education*, *13*(2), 5-25.

Koopmans, M. (2016). Investigating the Long Memory Process in Daily High School Attendance Data. In M. Koopmans & D. Stamovlasis (Eds.), *Complex Dynamical Systems in Education: Concepts, Methods and Applications* (pp. 299–321). Springer International Publishing.

Koopmans, M. (2020). Education is a complex dynamical system: Challenges for research. *The Journal of Experimental Education*, *88*(3), 358-374. https://doi.org/10.1080/00220973.2019.1566199

Labaree, D. (1992). Power, knowledge, and the rationalization of teaching: A genealogy of the movement to professionalize teaching. *Harvard educational review*, *62*(2), 123-155. https://doi.org/10.17763/haer.62.2.h73x7422v3166102

Langley, G. J., Moen, R. D., Nolan, K. M., Nolan, T. W., Norman, C. L., & Provost, L. P. (2009). *The Improvement Guide: A Practical Approach to Enhancing Organizational Performance*. John Wiley & Sons.

Lee Swanson, H., & Sachse-Lee, C. (2000). A Meta-Analysis of Single-Subject-Design Intervention Research for Students with LD. *Journal of Learning Disabilities*, *33*(2), 114–136.

https://doi.org/10.1177/002221940003300201

Lipsey, M. W. (1993). Theory as method: small theories of treatments. *New directions for program evaluation*, *57*, 5-38. https://doi.org/10.1002/ev.1637

Lipsky, M. (2010). *Street-Level Bureaucracy, 30th Ann. Ed.: Dilemmas of the Individual in Public Service*. Russell Sage Foundation.

Lortie, D. C. (1975). *Schoolteacher: A Sociological Study*. University of Chicago Press.

Matthes, N., Ogunbo, S., Pennington, G., Wood, N., Hart, M. K., & Hart, R. F. (2007). Statistical process control for hospitals: methodology, user education, and challenges. *Quality Management in Healthcare*, *16*(3), 205-214. https://doi.org/10.1097/01.QMH.0000281056.15177.a2

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*(2), 295-314. https://doi.org/10.1007/s10648-014-9287-x

Means, B., Chen, E., DeBarger, A., & Padilla, C. (2011). Teachers' Ability to Use Data to Inform Instruction: Challenges and Supports. *Office of Planning, Evaluation and Policy*

Development, US Department of Education.

Mehta, J. (2015). *The Allure of Order: High Hopes, Dashed Expectations, and the Troubled Quest to Remake American Schooling*. Oxford University Press.

Mertens, K., Vaesen, I., Löffel, J., Kemps, B., Kamers, B., Zoons, J., Darius, P., Decuypere, E., De Baerdemaeker, J., & De Ketelaere, B. (2009). An intelligent control chart for monitoring of autocorrelated egg production process data based on a synergistic control strategy. *Computers and Electronics in Agriculture*, *69*(1), 100–111. https://doi.org/10.1016/j.compag.2009.07.012

Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American journal of sociology*, *83*(2), 340-363. https://doi.org/10.1086/226550

Mohammed, M. A. (2004). Using statistical process control to improve the quality of health care. *Quality & Safety in Health Care*, *13*(4), 243–245. http://doi.org/10.1136/qshc.2004.011650

Montgomery, D. C. (2007). *Introduction to statistical quality control*. John Wiley & Sons.

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and Causal Inference*. Cambridge University Press.

Murnane, R. J., & Willett, J. B. (2010). *Methods Matter: Improving Causal Inference in Educational and Social Science Research*. Oxford University Press.

National Center on Response to Intervention. (2010). Users guide to universal screening tools chart. Washington, DC: National Center on Response to Intervention, Office of Special Education Programs, U.S. Department of Education.

National Research Council. (2002). *Scientific research in education* (L. Towne & R. J. Shavelson (eds.)). National Academy Press.

NYC Department of Education. (2018, September 10). *2015-2017 Historical Daily Attendance By School*. NYC Open Data. https://data.cityofnewyork.us/Education/2015-2017-Historical-Daily-Attendance-By-School/46g3-savk

Page, E. S. (1954). Continuous Inspection Schemes. *Biometrika*, *41*(1/2), 100–115. https://doi.org/10.2307/2333009

Payne, C. M. (2008). *So Much Reform, So Little Change: The Persistence of Failure in Urban Schools*. Harvard Education Press.

Payne, C. M., & Ortiz, C. M. (2017). Doing the Impossible: The Limits of Schooling, the Power of Poverty. *The Annals of the American Academy of Political and Social Science*, *673*(1), 32–59. https://doi.org/10.1177/0002716217719019

Perla, R. J., Provost, L. P., & Murray, S. K. (2011). The run chart: a simple analytical tool for learning from variation in healthcare processes. *BMJ Quality & Safety*, *20*(1), 46–51. http://dx.doi.org/10.1136/bmjqs.2009.037895

Provost, L. P., & Murray, S. (2011). *The Health Care Data Guide: Learning from Data for Improvement*. John Wiley & Sons.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. SAGE.

Reed, J. E., & Card, A. J. (2016). The problem with Plan-Do-Study-Act cycles. *BMJ Quality & Safety*, *25*(3), 147–152. http://dx.doi.org/10.1136/bmjqs-2015-005076

Research!America. (Fall 2019). 2013-2018 U.S. Investments in Medical and Health Research and Development. https://www.researchamerica.org/sites/default/files/Publications/InvestmentReport2019_Fnl.pdf

Results for America. (2018). ESSA Leverage Points: 50-State Report on Promising Practices for Using Evidence to Improve Student Outcomes. Results for America.

Roberts, S. W. (1959). Control Chart Tests Based on Geometric Moving Averages. *Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences*, *42*(1), 239–250.

Roderick, M. (2012). Drowning in Data but Thirsty for Analysis. *Teachers College Record*, *114*(11).

Rodriguez, M. C., & Nickodem, K. (2018, April). Comprehensive partitioning of student achievement variance to inform equitable policy design. *Paper Presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY*. https://conservancy.umn.edu/bitstream/handle/11299/195229/2018-

PartitioningVariance-NCME.pdf?sequence=1

Salanova, M., Llorens, S., García-Renedo, M., Burriel, R., BresÓ, E., & Schaufeli, W. B. (2005). Towards a Four-Dimensional Model of Burnout: A Multigroup Factor-Analytic Study Including Depersonalization and Cynicism. *Educational and Psychological Measurement*, *65*(5), 807–819. https://doi.org/10.1177/0013164405275662

Schilling, M. F. (2012). The Surprising Predictability of Long Runs. *Mathematics Magazine*, *85*(2), 141–149. https://doi.org/10.4169/math.mag.85.2.141

Schön, D.A. (1992). The Reflective Practitioner: How Professionals Think in Action (1st ed.). Routledge. https://doi.org/10.4324/9781315237473

Schouten, L. M., Hulscher, M. E., van Everdingen, J. J., Huijsman, R., & Grol, R. P. (2008). Evidence for the impact of quality improvement collaboratives: systematic review. *Bmj*, *336*(7659), 1491-1494. https://doi.org/10.1136/bmj.39570.749884.BE

Scott, S. L., & Varian, H. R. (2014). Predicting the present with Bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, *5*(1-2), 4-23. https://doi.org/10.1504/IJMMNO.2014.059942

Scrucca, L. (2004). qcc: An R package for quality control charting and statistical process control. *R News*, *4/1*, 11–17.

Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*. ASQ Quality Press.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational researcher*, *31*(7), 15-21. https://doi.org/10.3102/0013189X031007015

Slavin, R. E. (2020). How evidence-based reform will transform research and practice in education. *Educational Psychologist*, *55*(1), 21-31. https://doi.org/10.1080/00461520.2019.1611432

Speroff, T., & O'Connor, G. T. (2004). Study designs for PDSA quality improvement research. *Quality Management in Healthcare*, *13*(1), 17-32.

Stokes, D. E. (2011). *Pasteur's Quadrant: Basic Science and Technological Innovation*. Brookings Institution Press.

Stoumbos, Z. G., & Reynolds, M. R. (2000). Robustness to non-normality and autocorrelation of individuals control charts. *Journal of Statistical Computation and Simulation*, *66*(2), 145–

187. https://doi.org/10.1080/00949650008812019

Stoumbos, Z. G., Reynolds, M. R., Ryan, T. P., & Woodall, W. H. (2000). The State of Statistical Process Control as We Proceed into the 21st Century. *Journal of the American Statistical Association*, *95*(451), 992–998. https://doi.org/10.2307/2669484

Superville, C. R., & Adams, B. M. (1994). An evaluation of forecast-based quality control schemes. *Communications in Statistics-Simulation and computation*, *23*(3), 645-661. https://doi.org/10.1080/03610919408813191

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285–1293. https://doi.org/10.1126/science.3287615

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological science in the public interest*, *1*(1), 1-26. https://doi.org/10.1111/1529-1006.001

Taylor, M. J., McNicholas, C., Nicolay, C., Darzi, A., Bell, D., & Reed, J. E. (2014). Systematic review of the application of the plan–do–study–act method to improve quality in healthcare. *BMJ quality & safety*, *23*(4), 290-298. http://doi.org/10.1136/bmjqs-2013-001862

Tyack, D. B., & Cuban, L. (1995). *Tinkering Toward Utopia*. Harvard University Press.

United States. (1978). *The Belmont Report. Ethical principles and guidelines for the protection of human subjects of research*. The Commission.

Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative science quarterly*, 1-19. https://doi.org/10.2307/2391875

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, *10*(4), 843-876. https://doi.org/10.1080/19345747.2017.1300719

What Works Clearinghouse. (2020). *What Works Clearinghouse Standards Handbook, Version 4.1*. Washington, DC: U.S. Department of Education, Institute of Educational Services, National Center for Education Evaluation and Regional Assistance. This report is available on the What Works Clearinghouse website at https://ies.ed.gov/ncee/wwc/handbooks.

White, C. M., Statile, A. M., Conway, P. H., Schoettker, P. J., Solan, L. G., Unaka, N. I., Vidwan, N., Warrick, S. D., Yau, C., & Connelly, B. L. (2012). Utilizing improvement science methods to improve physician compliance with proper hand hygiene. *Pediatrics*, *129*(4), e1042–

e1050. https://doi.org/10.1542/peds.2011-1864

Wiemken, T. L., Furmanek, S. P., Mattingly, W. A., Wright, M.-O., Persaud, A. K., Guinn, B. E., Carrico, R. M., Arnold, F. W., & Ramirez, J. A. (2018). Methods for computational disease surveillance in infection prevention and control: Statistical process control versus Twitter's anomaly and breakout detection algorithms. *American journal of infection control*, *46*(2), 124-132. https://doi.org/10.1016/j.ajic.2017.08.005

Woodall, W. H. (2000). Controversies and Contradictions in Statistical Process Control. *Journal of Commodity Science, Technology and Quality*, *32*(4), 341–350. https://doi.org/10.1080/00224065.2000.11980013

Woodall, W. H. (2017). Bridging the gap between theory and practice in basic statistical process monitoring. *Quality Engineering*, *29*(1), 2–15. https://doi.org/10.1080/08982112.2016.1210449

Woodall, W. H., & Montgomery, D. C. (2014). Some current directions in the theory and application of statistical process monitoring. *Journal of Quality Technology*, *46*(1), 78-94. https://doi.org/10.1080/00224065.2014.11917955

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Nelson Education.

Yamada, H., & Bryk, A. S. (2016). Assessing the first two years' effectiveness of Statway®: A multilevel model with propensity score matching. *Community College Review*, *44*(3), 179–204.

Yeager, D. S., Bryk, A. S., Muhich, J., Hausman, H., & Morales, L. (2013). Practical measurement. *Palo Alto, CA: Carnegie Foundation for the Advancement of Teaching*. https://labs.la.utexas.edu/adrg/files/2013/12/Practical-Measurement.pdf

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of educational Research*, *81*(2), 267-301. https://doi.org/10.3102/0034654311405999

Yurkofsky, M. M., Peterson, A. J., Mehta, J. D., Horwitz-Willis, R., & Frumin, K. M. (2020). Research on Continuous Improvement: Exploring the Complexities of Managing Educational Change. *Review of Research in Education*, *44*(1), 403–433. https://doi.org/10.3102/0091732X20907363

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*(4), 561–577. https://doi.org/10.1093/clinchem/39.4.561